DISS. ETH NO. 25428

PHYLOGENETIC COMPARATIVE METHODS IN THE ERA OF BIG DATA

A thesis submitted to attain the degree of
DOCTOR OF SCIENCES of ETH ZURICH
(Dr. sc. ETH Zurich)

presented by

VENELIN MITOV
M.Sc. ETH Zurich, Zurich, Switzerland
born on 20.08.1981
living in Basel, BS, citizen of Bulgaria

accepted on the recommendation of

Prof. Dr. Tanja Stadler, examiner
Prof. Dr. Roland Robert Regös, co-examiner
Prof. Dr. Samuel Alizon, co-examiner

2018

ACKNOWLEDGMENTS

I wish to thank all my scientific colleagues who helped me through my doctoral studies. My supervisor, Tanja Stadler, has been the most inspiring and supportive person. Work with her has resulted in Chapters 2, 3, 5, 6 and 7. For Chapter 2, I am indebted to all colleagues from the cEvo group who took part in organizing the "Taming the Beast" summer school in phylogenetics and, in particular, to Louis du Plessis, Joëlle Barido-Sottani and Veronika Bošková, who invested the most effort in writing our co-authored publication. I am very grateful to Roland Regoës for inviting me to participate in the study of HIV virulence conducted in his research group. This study has resulted in the co-authored Chapter 4. I appreciate a lot the contribution from Krzysztof Bartoszek who shared his mathematical expertise and actively took part in two of the projects resulting in Chapters 6 and 7. I also wish to acknowledge the contribution from Georgios Asimomitis who worked on the PCMBase R-package during his three-week lab-rotation in the cEvo group. I wish to thank Emma Hodcroft and Andrew Leigh-Brown for sharing the anonymized HIV data used for the analysis in Chapter 3 and Jörg Stelling for getting Tanja and me interested in the brain- and body-mass allometry in mammal species which has become the main biological example in Chapter 7. I am grateful to all my other co-authors: Denise Kühnert, Carsten Magnus, Nicola Müller, Jūlija Pečerska, David Rasmussen, Chi Zhang, Alexei Drummond, Tracy Heath, Oliver Pybus, Timothy Vaughan, Frederic Bertels, Alex Maryel, Gabriel Leventhal, Jacques Fellay, Huldrych Günthard, Jürg Boni, Sabine Yerli, Thomas Klimkait, Vincent Aubert, Manuel Battegay, Andri Rauch, Mattias Cavassini, Alexandra Calmy, Enos Bernasconi, Patrick Schmid, Alexandra Scherrer, Viktor Müller, Sebastian Bonhoeffer and Roger Kouyos. I would like to thank Joëlle Barido-Sottani one more time for helping me with the translation of the summary of this thesis in grammatically correct French. Finally, I am very grateful to Samuel Alizon and Roland Regoës for accepting to be co-examiners of my work.

Plenty of people were not directly involved in producing the contents of my thesis but deserve my heartfelt gratitude just as well. Brigitta Elhadj-Keller, Sibylle Meneghetti, Andrea Huber, Cindy Malnasi for their swift help with organizational and administrative matters. All my officemates during the past four and a half years – the time spent with you was enjoyable and enriching for me on a personal level, despite the sometimes long hours of silence when typing on our keyboards was the only sound to hear. All of the members of the cEvo group for the daily lunches and monthly sushi lunches when we gathered in the Science lounge at D-BSSE and had so many moments of unforgettable fun. Mariola Bozova, Christina Zeller, Claudia Pleuss and her family, Rachel and Graham Warnock, Hesam Montazeri, David Rasmussen, Enkelejda Miho, Laura Caroll, Simone Tobler, Tamara Hell, Dimitar Nemski and Dimitar Shiyachki – I am glad to call you my friends.

Finally, I thank my family for being my people. Dear Mom and Dad, thank you for having raised me into what I am today and for supporting me through all my life.

CONTENTS

Part I

PREFACE

SUMMARY

Phylogenetic comparative methods (PCMs) are used for studying the evolution of various biological species, ranging from micro-organisms to animals and plants. These methods are based on the computer-assisted comparison of phenotype and molecular sequence data in populations of living and/or extinct species or organisms. With the rise of genome sequencing, it has become possible to infer the phylogenetic trees of large populations, such as the entire mammal clade, counting nearly 4000 species, or the transmission trees from large epidemic outbreaks, counting thousands to hundreds of thousands of infections. This has encouraged the transfer of PCMs developed originally for studying a few quantitative traits in a small phylogeny of living species to data much bigger in size and, sometimes, different in type.

In this thesis, I explore several difficulties encountered in the application of PCMs for the study of big phylogenetically linked comparative data. These range from technical problems, such as the development of fast algorithms for phylogenetic model inference to conceptual issues, such as the difference between an epidemic and a population of sexually reproducing organisms in estimating the heritability of a quantitative trait. My approach is a mixture of a top-down and a bottom-up strategy. At the high level, I start from poorly understood biological questions, for which comparative data has been available, and I identify particular issues hindering the use of existing PCMs to analyse that data. Then, I develop a prototype solving these issues for the data in question. Finally, I consider the prototype in a broader perspective, searching for possibilities to apply the same solution to a more general class of problems, without compromising the computational efficiency. This approach led to the development of several software tools, which, I hope, would prove useful in future studies.

The first chapter gives a general historical background and introduces the main concepts of PCMs. The rest of the thesis is divided in two parts. The part "Publications" (Chapters 2, 3 and 4) includes articles published during my doctoral studies. Chapter 2 introduces the field of phylogenetics and the software tools used for inferring phylogenetic trees based on molecular sequence data. Phylogenetic trees of that kind represent the main input for all methods described in the following chapters. In Chapter 3, I study the effects of within-host pathogen evolution on various estimators of the set-point viral load heritability in HIV patients. Based on simulations and real data of nearly ten thousand HIV patients, I show that neglecting or inaccurately accounting for within-host pathogen evolution has been the main cause for a long-standing discrepancy between different estimates of the set-point viral load heritability. Chapter 4 makes use of these results to estimate the heritability of two additional HIV traits: the CD4 cell decline and the per-parasite pathogenicity. The part "Manuscripts" (Chapters 5, 6 and 7) includes works which, at the time of submitting this thesis, are in revision or in preparation for submission to peer-reviewed journals. In Chapter 5, I develop generic algorithms for parallel traversal of phylogenetic trees, with "traversal" meaning the application of an abstract operation to all nodes in the tree, while respecting their hierarchical order. I implement these algorithms within the C++ library SPLITT intended as a fast back-end for higher level packages, such as generic PCM implementations. Chapter 6 describes one such tool – the R-package PCMBase implementing fast likelihood calculation of multi-

trait Gaussian phylogenetic models. The poor efficiency of the likelihood calculation is the principal bottleneck in applying Gaussian phylogenetic models to big phylogenetic trees. PCMBase resolves this issue for a very large family of models and all types of phylogenetic trees, including non-ultrametric trees and polytomies. Taking advantage of PCMBase and SPLITT, in Chapter 7, I analyze the biggest published phylogeny of mammal species, for which brain and body mass measurements are available. Based on this data, I show that present-day PCMs are unable to model the heterogeneity of the evolutionary process across different mammal clades. As a solution, I propose a new method for inferring jointly a set of different evolutionary models on different parts of the tree.

Finally, in Chapter 8, I discuss the methods developed in this thesis and suggest directions for future research.

RÉSUMÉ

Les méthodes phylogénétiques comparatives (MPC) sont appliquées dans l'étude de l'évolution de plusieurs espèces biologiques, allant de micro-organismes aux espèces animales et végétales. Ces méthodes sont basées sur la comparaison assistée par ordinateur de phénotypes et de séquences moléculaires dans des populations d'organismes ou d'espèces vivantes aussi bien que disparues. Avec le développement du séquençage de génomes, il est devenu possible de déduire les arbres phylogénétiques de grandes populations, comme le clade complet des mammifères, qui compte près de 4000 espèces ou les arbres de transmission de grandes épidémies comptant des milliers à des centaines de milliers d'infections. Cela a encouragé le transfert de MPC développées à l'origine pour étudier quelques caractères quantitatifs dans une petite phylogénie d'espèces vivantes à des données de taille beaucoup plus grande et parfois d'un genre différent.

Dans cette thèse, j'explore plusieurs difficultés rencontrées dans l'application de MPC à des données comparatives de grande taille. Celles-ci vont de problèmes techniques, tels que le développement d'algorithmes rapides pour l'inférence de modèles phylogénétiques aux problèmes conceptuels, tels que la différence entre une épidémie et une population d'organismes se reproduisant sexuellement lors de l'estimation de l'héritabilité d'un caractère quantitatif. Mon approche est un mélange de stratégie descendante et de stratégie ascendante. Au niveau supérieur, je commence par des questions biologiques mal comprises, pour lesquelles des données comparatives sont disponibles, et j'identifie des problèmes particuliers qui entravent l'utilisation des MPC existants pour l'analyse de ces données. Ensuite, je développe une solution prototype spécialisée pour les données en question. Enfin, je considère le prototype dans une perspective plus large, en étudiant la possibilité d'appliquer la même solution à une classe de problèmes plus générale, sans compromettre l'efficacité du calcul. Cette approche a abouti au développement de plusieurs logiciels, qui, je l'espère, pourraient s'avérer utiles dans de futures études.

Le premier chapitre présente le contexte historique général et les concepts de base des MPC. Le reste de la thèse est divisé en deux parties. La partie «Publications» (chapitres 2, 3 et 4) comprend des articles publiés au cours du doctorat. Le chapitre 2 présente le domaine de la phylogénétique et les outils logiciels utilisés pour déduire des arbres phylogénétiques à partir de séquences moléculaires. Les arbres phylogénétiques de ce type représentent les données d'entrée principales pour toutes les méthodes décrites dans les chapitres suivants. Dans le chapitre 3, j'étudie les effets de l'évolution des pathogènes au sein de l'hôte sur plusieurs estimateurs de l'héritabilité de la concentration virale dans le sang des patients atteints du VIH. Sur la base de simulations et de données réelles issues de près de dix mille patients infectés par le VIH, j'ai montré que négliger ou modéliser de manière inexacte l'évolution intra-hôte des pathogènes était la principale cause d'une divergence de longue date entre différentes estimations de l'héritabilité de la concentration virale. Le chapitre 4 utilise ces résultats pour estimer l'héritabilité de deux caractéristiques supplémentaires du VIH: le déclin des cellules CD4 et la pathogénicité par parasite. La partie «Manuscrits» (chapitres 5, 6 et 7) comprend des travaux qui, au moment de la soumission de cette thèse, sont en révue, en révision ou en préparation pour être soumis à des journaux scientifiques. Dans le chapitre 5, je développe des algorithmes génériques pour la traversée parallèle d'arbres phylogéné-

tiques, où «traverser» signifie appliquer une opération abstraite à tous les nœuds de l'arbre, tout en respectant leur ordre hiérarchique. Je mets en œuvre ces algorithmes au sein de la bibliothèque C++ SPLITT conçue comme un back-end rapide pour des packages plus complexes, tels que des implémentations de MPC génériques. Le chapitre 6 décrit un de ces outils - le package PCMBase écrit en R, qui met en œuvre un calcul de probabilité rapide de modèles phylogénétiques Gaussiens à caractères multiples. La faible efficacité du calcul de cette probabilité est l'obstacle principal à l'application de modèles phylogénétiques Gaussiens aux grands arbres phylogénétiques. PCMBase résout ce problème pour une très grande famille de modèles et tous les types d'arbres phylogénétiques, y compris les arbres et les polytomies non ultramétriques. En tirant parti de PCMBase et de SPLITT, au chapitre 7, j'analyse la plus grande phylogénie publiée sur les espèces de mammifères, pour laquelle des mesures de masse cérébrale et corporelle sont disponibles. Sur la base de ces données, je montre que les MPC actuelles sont incapables de modéliser l'hétérogénéité du processus évolutif à travers différents clades de mammifères. Pour résoudre ce problème, je propose une nouvelle méthode pour inférer conjointement un ensemble de modèles évolutifs différents sur des parties différentes de l'arbre.

Enfin, au chapitre 8, je discute les méthodes développées dans cette thèse et propose des orientations pour les recherches futures.

1

INTRODUCTION

Species that diverged recently on the tree of life are likely to be phenotypically similar. This phenomenon known as "phylogenetic effect" undermines the classical statistical methods applied to comparative inter-species data, because they ignore the species' shared history (Felsenstein, 1985). Felsenstein (1985) was the first to propose a straightforward solution – his famous method of phylogenetic independent contrasts. This event marked the birth of a new field in evolutionary biology – the phylogenetic comparative methods (PCMs).

Who could have imagined the immense expansion of PCMs through the next decades (Pennell and Harmon, 2013)? From a nuisance that has to be cleaned from the comparative data prior to statistical test, the phylogenetic effect is nowadays regarded as a fundamental source of information about the past evolution of the species (Losos, 2011). To some extent this progress owes to two interacting lines of development:

- the innovation of fast genetic sequencing techniques providing novel sequence data from an ever wider range of organisms;
- the development of fast phylogenetic inference tools capable to provide time calibrated trees from multiple sequence alignments going beyond 10'000 sequences (Price, Dehal, and Arkin, 2009; Stamatakis, Hoover, and Rougemont, 2008).

Big species trees, counting thousands of species have become available (see, e.g. (Bininda-Emonds et al., 2007)). Moreover, the sequencing of rapidly evolving micro-organisms, such as RNA viruses collected from patients during epidemics, provides unprecedented amounts of sequences, based on which it is possible to infer approximate transmission trees for thousands of patients. Many PCMs invented originally for the analysis of macro-evolutionary data from living species have been adopted by epidemiologists for the study of pathogen traits, such as the virulence of human immunodeficiency virus (HIV) and malaria infections (Alizon et al., 2010; Anderson et al., 2010; Hodcroft et al., 2014; Shirreff et al., 2013).

However, the transfer of PCMs from macro-evolutionary to epidemiological types of data as well as the transfer from small to big data size hides both conceptual and technical challenges. In this thesis, I consider in detail two of these challenges:

1. The theoretical basis of PCMs is the quantitative genetics theory of sexually reproducing species. Since infections represent clonal transmissions of pathogens between hosts, the mechanisms of sexual reproduction are not present and a straightforward transfer of the theory is not possible. The differences between pathogens and mating species need to be identified and fundamental concepts, such as the definition of heritability of a pathogen trait need a thorough rethinking.
2. Trees inferred from large populations of species or global scale epidemics are characterized by heterogeneous evolutionary processes varying in rates of evolution and selective pressures. Few of the existing PCMs can infer such heterogeneities and if they do, these are limited to small ultrametric trees (all species sampled at the current time). Extending these methods to support large non-ultrametric transmission trees is both a

modeling and a technical challenge. The existing models must be extended to fit the varying patterns of phenotypic correlation in different parts of the tree. The big size of the trees necessitates fast algorithms for fitting such complex models.

In this chapter, I introduce the basics of PCMs. As Harmon (2018) wrote PCMs "... stem from and bring together three main fields: population and quantitative genetics, paleontology, and phylogenetics". Of the above three fields, only a basic knowledge of quantitative genetics is indispensable to understand the following chapters in the thesis. I will not discuss paleontology, because I did not have the chance to analyze paleontological data during the thesis. The field of phylogenetics is relevant for my work up to the point of using tools for phylogenetic inference. I discuss this topic briefly at the end of this chapter, while the co-authored Chapter 2 of the thesis represents a general introduction to the field with focus on tools for Bayesian phylogenetic inference.

1.1 A BRIEF OVERVIEW OF QUANTITATIVE GENETICS

The main objects of study in both population and quantitative genetics are one to several consecutive generations in a population of organisms. The difference is that, while population genetics focus on properties of genes, e.g. the change in allele frequencies from one generation to the next due to genetic mutation and selection, quantitative genetics estimate properties, e.g. the heritability, of continuous characters, also known as quantitative traits. Quantitative traits represent characters that can be measured in real numbers, such as body mass or systolic pressure. I will use the terms "trait measurement" or "trait value" to denote the measured value of a trait for a given individual in a population. Conversely, I use the term "phenotypic value" to denote the true value, excluding possible measurement error.

1.1.1 *Quantitative trait loci*

A principal goal of quantitative genetics is to partition the phenotypic variance in a population into components attributable to genetic and environmental factors. Fundamental for the study of the genetic and environmental sources of variance is the general linear model for the phenotype (Lynch and Walsh (1998), ch. 6), in which, for a given trait of interest, the observed phenotypic value, z, of an organism is represented as a sum of effects of the organism's genes, G, general (macro-) environmental effects, E, gene by (macro-) environment interaction, I and special (micro-) environmental effects, e:

$$z = G + I + E + e \tag{1.1}$$

It is assumed that the trait is influenced by a number of genes whose locations in the species' genome are called quantitative trait loci (QTLs). In an individual, the configuration of gene-variants (alleles) found at the trait's QTLs is called genotype and, for a population, the genotypic value, $G_{\mathbf{x}}$, of a genotype $\mathbf{x}$ is defined as the expected phenotypic value of its carriers: $G_{\mathbf{x}} = \mathrm{E}(z|\text{genotype} = \mathbf{x})$. The remaining terms in eq. 1.1 are "defined in a least-squares sense as deviations from lower order expectations" (Lynch and Walsh, 1998). It is worthy to note that $G_{\mathbf{x}}$ depends on the distribution of $\mathbf{x}$ across environments in the population and that, by construction, the residuals $z - G = I + E + e$ have zero mean and are uncorrelated with G (Lynch and Walsh (1998), ch. 6). Thus, the total phenotypic variance

observed in the population can be partitioned into a component that is purely genetic and a component that is attributable to both, non-genetic (purely environmental) factors as well as gene-by-environment interactions: $Var(z) = Var(G) + Var(z - G)$.

1.1.2 *The heritability of quantitative traits*

Evolution of a trait under mutation and selection is only possible if the trait is "heritable". In simple terms, the term "heritability" summarizes the relationship between genes and phenotypes. But can a single word speak for all possible manifestations and consequences of this relationship? Jacquard (1983) noticed that heritability has been used by quantitative geneticists to serve (at least) three different concepts: (i) the genetic determination of a trait; (ii) the resemblance between relatives; (iii) the efficiency of selection. Hence, it is often confusing to use the term "heritability" without an accompanying definition or a qualifier like "narrow-sense", "broad-sense" and "realized". This distinction of concepts has become somewhat vague in subsequent works. For instance, the heritability in the narrow-sense, which is essentially a lower bound for the genetic determination of a trait, has often been considered equivalent up to a known scaling factor to the parent-offspring regression slope, which is used to measure the resemblance between family members (Lynch and Walsh, 1998). This might not present an issue in the study of sexually reproducing populations, where the above equivalences have been well studied. But when transferring this theory to a different domain, such as the evolution of pathogens, a confusion arises. I think that the confusion of terms like narrow- and broad- sense heritability has been a major obstacle causing a long-lasting discrepancy between various studies of the heritability of HIV virulence. Since this discrepancy is the topic of Chapters 3 and 4 of this thesis, I introduce the above concepts and terminology. In chapter 3, I'll overview the same concepts from the point of view pathogen traits.

GENETIC DETERMINATION. Considering a quantitative trait, the degree to which the genes of individuals determine their phenotypic values is quantified in a statistical sense by the **broad-sense heritability**, H^2. H^2 is defined as the ratio of the variance of genotypic values to total phenotypic variance in the population (Falconer, 1996):

$$H^2 = Var(G)/Var(z) \tag{1.2}$$

.

Assuming a sufficiently large population and full knowledge of the distinct genotypes influencing the trait, H^2 can be measured by the coefficient of determination, R^2_{adj}, estimated over a grouping of the population by genotype. In the world of animals and plants, though, it is practically impossible to measure H^2 in this way, because population sizes are small compared to large numbers of (usually unknown) genotypes. Thus, quantitative genetics focuses on estimating a lower bound for H^2 – the **narrow-sense heritability**, h^2. h^2 summarizes how much of the trait variance is attributable to single-locus additive genetic effects. A formal definition of additive genetic effect is overly technical, so I leave it beyond the scope of this introduction (it can be found in (Lynch and Walsh, 1998)). Nevertheless, it is important to note that, in sexually reproducing populations, h^2 can be estimated from measures of the trait-resemblance between relatives.

RESEMBLANCE BETWEEN RELATIVES. Relatives resemble each other not only for carrying similar genes but also for living in similar environments. Hence, it is necessary to disentangle the concept of resemblance from that of genetic determination. For an ordered relationship such as parent-offspring, the resemblance is usually measured by the **regression slope**, b, of expected offspring values on mean parental values. For members of unordered relationships, such as identical twins, sibs and cousins, their relative resemblance is quantified by the one-way analysis of variance (ANOVA), which estimates the so-called **intraclass correlation** (ICC) denoted here as $r_A[\text{type of relationship}]$.

EFFICIENCY OF SELECTION. The last of the three concepts identified by Jacquard (1983) is that of the efficiency of selection for breeding of the individuals with "best" trait-values. In breeding experiments the goal is to optimize a trait by repetitive artificial selection for reproduction of the "best" individuals in a generation. A textbook example is truncation selection in which only individuals with measurements above a given threshold are allowed to reproduce. For a generation, the difference $\Delta_s = \mu_s - \mu$ between the mean value of individuals selected for reproduction, μ_s, and the mean of the generation, μ, is called the selection differential. Denoting by the mean of the offspring generation, the difference $R = \mu_o - \mu$, is called the response to selection. Then, the efficiency of the truncation selection is measured by the **realized heritability** (Hartl and Clark, 2007), defined as the ratio:

$$h^2_R = R/\Delta_s \tag{1.3}$$

CONNECTING THE DOTS. The success of quantitative genetics in the pre-genomic era relies on the insight that "*inferences concerning the genetic basis of quantitative traits can be extracted from phenotypic measures of the resemblance between relatives* (Lynch and Walsh, 1998)". Mathematically, this quote is expressed as a set of approximations, which have become dogmatic in quantitative genetics:

$$H^2 = R^2_{adj} \simeq r_A[\text{identical twins}]$$

$$h^2 \simeq b \simeq 4r_A[\text{half sibs}] \simeq h^2_R.$$

The first equation is valid in general, provided there is no strong maternal effect on the trait, the observed twins have been separated at birth and raised in independent environments and the assumptions of ANOVA such as normality and homoscedasticity are at least approximately met. The second equation, though, is provable only for diploid sexually reproducing species. This is because genetic segregation and recombination during sexual reproduction ensure that single-locus additive effects are inherited at bigger proportions (1/2 from each parent) compared to multi-locus (epistatic) interactions (i.e. 1/4 for 2-loci-, 1/8 for 3-loci-interactions, etc) (Falconer, 1996; Lynch and Walsh, 1998).

In summary, quantitative genetics deals with the properties of quantitative traits, observable over one to several generations of a population. In sexually reproducing populations, heritability is used to quantify to what extent the genetics explain a trait (broad-sense heritability, H^2) as well as to measure or predict the response to trait-based selection for reproduction (realized heritability, h^2_R). Since it is practically hard to measure H^2, one often uses empirical measures of the resemblance between relatives (i.e. parent-offspring regression, b, or ICC from half sibs, r_A) to estimate the extent, to which single-locus additive effects de-

termine the trait (narrow-sense heritability, h^2). It turns out that $h^2 \simeq h_R^2$, justifying the dual role of h^2 as a measure of genetic determination and a measure for the rate of trait-evolution resulting from selection.

1.2 MODELING RANDOM GENETIC DRIFT AND SELECTION OVER LONG PERIODS OF TIME

Population and quantitative genetics provide useful models of character evolution in a population on the scale of one to several generations. In contrast, macro-evolutionary studies focus on the evolution of species that can only be observed over long periods, ranging from hundreds to millions of generations. Simpson, 1953 used models from population genetics to interpret the observed evolution of teeth in the fossil record of horses. In analogy with the idea of adaptive topographies for genotypes proposed by Wright, 1931, Simpson proposed the concept of adaptive zones for phenotypes. Two decades later, Lande, 1976 provided a mathematical interpretation of Simpson's adaptive zones, which is equivalent to the contemporary notion of a "fitness landscape": the (planar) dimensions of the landscape represent the mean phenotypic values of a set of quantitative traits measured in a population, while the vertical dimension (i.e. the height of the landscape) is the population's mean fitness.

Based on a simple equation from quantitative genetics – the definition of realized heritability (eq. 1.3) – Lande, 1976 derived a quantitative formula for the expected mean phenotype in a population evolving under random drift and selection for an arbitrary period of time. Decades later, this result became the theoretical basis of numerous phylogenetic models of quantitative trait evolution including the ones discussed in this thesis. Hence, I briefly outline the assumptions and the resulting expressions following the derivation in Lande, 1976.

It is assumed that a population of effective size N evolves as a sequence of generations. Each generation undergoes the following order of events: reproduction (birth), selection and random sampling. As Lande, 1976 summarizes, "N individuals are drawn at random from the selected population to constitute the parents of the next generation". A quantitative trait z is measured as the difference from some optimum value. It is assumed that the selection for reproduction is stabilizing around this optimum and is described by a Gaussian (normal) fitness function:

$$\Phi(z) = \exp\left(\frac{z^2}{2w^2}\right), \tag{1.4}$$

where w denotes the width of the adaptive zone. At any time t, the distribution of z in the current population is assumed to be normal with mean $\bar{z}(t)$ and variance σ_T^2, which is constant with respect to both t and $\bar{z}(t)$. Also, it is assumed that the realized heritability, h_R^2, is constant with respect to t and $\bar{z}(t)$ and that the selection is weak, that is $\sigma_T^2 \ll w^2$. Then, given an initial population mean $\bar{z}(0)$, for any time $t > 0$ the distribution of the population mean $\bar{z}_t$ is normal with mean and variance (Lande, 1976):

$$\begin{aligned} E\left[\bar{z}(t)\right] &= \bar{z}(0)\exp\left(-\tfrac{h_R^2\sigma_T^2}{w^2+\sigma_T^2}t\right) \\ Var\left[\bar{z}(t)\right] &= \tfrac{w^2+\sigma_T^2}{2N}\left[1-\exp\left(-2\tfrac{h_R^2\sigma_T^2}{w^2+\sigma_T^2}t\right)\right]. \end{aligned} \tag{1.5}$$

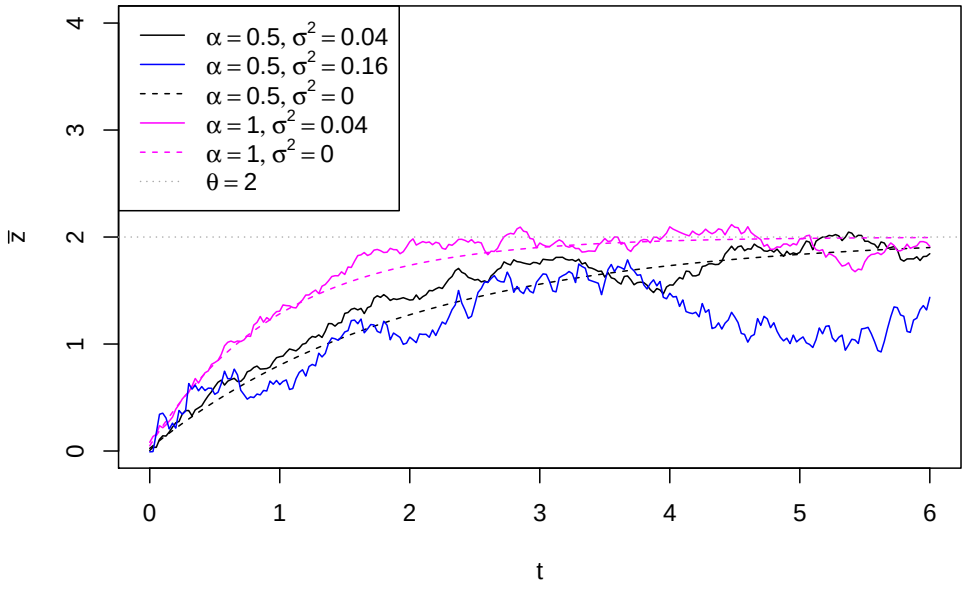


Figure 1.1: Random trajectories of OU processes with an optimum $\theta = 2$, initial value $\bar{z}(0) = 2$ and different values for the parameters α and σ^2.

Substituting $\theta = 0$, $\alpha = \frac{h_R^2 \sigma_T^2}{w^2 + \sigma_T^2}$, $\sigma^2 = \frac{h_R^2 \sigma_T^2}{N}$ in eq. 1.5 and denoting by $W(t)$ the standard Wiener process, we obtain the solution of an Ornstein-Uhlenbeck stochastic differential equation (SDE) with initial state $\bar{z}(0)$, long term optimum θ, strength α, and unit-time variance σ^2 (Lande, 1976; Uhlenbeck and Ornstein, 1930):

$$\begin{aligned} \text{OU: } d\bar{z}(t) &= \alpha\big[\theta - \bar{z}(0)\big]dt + \sigma dW(t) \\ \text{Solution:} & \\ E\big[\bar{z}(t)\big] &= \bar{z}(0)\exp(-\alpha t) \\ Var\big[\bar{z}(t)\big] &= \tfrac{\sigma^2}{2\alpha}\big[1 - \exp(-2\alpha t)\big]. \end{aligned} \tag{1.6}$$

Hence, the Ornstein-Uhlenbeck (OU) stochastic process defined in eq. 1.6 has been adopted by evolutionary biologists as a model of evolution under stabilizing selection and random drift around an optimum point in an adaptive zone of the fitness landscape. Figure 1.1 shows examples of random OU trajectories for different values of α and σ^2.

The parameter $\alpha > 0$ denotes the selection strength of the OU-process. In the limit $\alpha \to 0$, the OU process is equivalent to Brownian motion (BM) process with unit time variance increment σ^2. Seen as a model of neutral trait evolution under random genetic drift, the BM process has been the first model of evolution incorporated in PCM analysis of quantitative traits (Felsenstein, 1985). Later, Hansen introduced the OU model to the PCM field by proposing it as model for modeling evolution under stabilizing selection (Hansen, 1997). Since then, multiple extensions of these models have been proposed to accommodate various evolutionary concepts, such as adaptive radiation, punctuated equilibrium, directional selection, truncated selection and others (reviewd in (Pennell and Harmon, 2013)). In the next section I explain how the BM and the OU process are integrated in phylogenetic comparative methods.

1.3 PHYLOGENETIC TREES

The field of phylogenetics has been relevant for my work up to the point of using tools for fast inference of big phylogenetic trees from sequence alignments counting up to 10'000 sequences. Such phylogenetic trees constitute the main input for PCMs. The referenced articles (Price, Dehal, and Arkin, 2009; Stamatakis, Hoover, and Rougemont, 2008) provide comprehensive overview of maximum likelihood methods for inferring phylogenetic trees. Chapter 2 of this thesis provides an overview of tools for Bayesian inference of phylogenetic trees.

In this thesis I will be using three types of phylogenetic trees:

- transmission trees inferred from HIV genetic sequences extracted from infected patients. This type of phylogenetic tree represents the main input in Chapters 3 and 4.
- species trees inferred from the reference genetic sequences of living species. Such trees are called ultrametric to denote that all tips in the tree are located at the same time distance from the root. This type of tree will be the main input in Chapter 7.
- simulated birth-death trees representing simulated speciation and extinction histories. Such simulated trees are used for validating some of the implemented methods in Chapters 5, 6 and 7.

1.4 PHYLOGENETIC COMPARATIVE METHODS

In this thesis I focus on the subset of PCMs dealing with quantitative traits. In macro-evolutionary comparative analysis, the traits usually are the average values from measurements in finite populations representative of different biological species. In epidemiological studies, the measurements are taken from individual patients.

1.4.1 *Comparative data*

Comparative data constitutes one of the two inputs to a PCM. This is a set of N (possibly multi-)trait measurements from different species (in macro-evolutionary studies), or patients (in epidemiological studies). The other input of a PCM represents a rooted time-calibrated phylogenetic tree with N tips (species or patients), corresponding to the entries in the comparative data. In the case of multiple traits, I use the symbol k to denote the number of traits and $\vec{x}_i$ to denote the k-variate trait vector measured for individual i. I use the symbol $\mathbf{X}$ to denote the $k \times N$ matrix, the columns of which represent the trait vectors for the N individuals.

1.4.2 *Phylogenetic models of trait evolution*

At the heart of every PCM analysing the evolution of quantitative traits, there is a model of quantitative trait evolution. In the previous section I've introduced the BM and the OU models of trait evolution on the time scale of many generations of a finite population. Transferring these models to PCMs boils down to using their branching analogs. In essence, it is assumed that a k-variate parametric stochastic process starts from an initial k-vector at the root of the phylogenetic tree. At a branching point, the process splits into two processes, each

one inheriting the last state (trait value) at the branching point. For all developments in this thesis, it will be assumed that the two processes do not interact between each other after the split. While not valid in practice, this will be one of the key assumptions enabling fast model inference on big phylogenetic trees. Scenarios with interaction between the processes (i.e. co-evolution between forking lineages on the tree) have been considered in other works (see, e.g. Manceau, Lambert, and Morlon, 2016).

1.4.3 *Multivariate Gaussian distributions*

The key assumption in all PCMs studied throughout this thesis is that the joint distribution of all trait measurements associated with the tips in the phylogeny is multivariate Gaussian (also called multivariate normal). By definition, this is equivalent to the requirement that every linear combination $\vec{a}^T\vec{\mathbf{X}}$, where $\vec{a}$ is a column vector of kN real coefficients and $\vec{\mathbf{X}}$ is the kN-vector formed by concatenating all trait vectors in the comparative data, is a normally distributed random variable. The above implies that every single trait for every single species or individual is also a normal random variable. This follows simply from applying the rule to a vector $\vec{a}$, in which one coefficient is set to 1 and all others to 0.

The kN-variate Gaussian distribution is fully characterized by its mean vector, $\vec{\mu} \in \mathbb{R}^{kN}$, and its symmetric positive-definite variance-covariance matrix $\Sigma \in [\mathbb{R}]_{kN\times kN}$. Assuming that the phylogenetic model is one of the above mentioned branching stochastic processes (BM, OU or some of their variants), $\vec{\mu}$ and Σ are functions of the model parameters and the topology and branch lengths of the phylogenetic tree. Denoting the parameters of the model by θ and the phylogenetic tree by $\mathcal{T}$, the probability density function of the multivariate Gaussian distribution for a concatenated vector of trait values $\vec{\mathbf{X}}$ is given by

$$pdf(\vec{\mathbf{X}}|\theta,\mathcal{T}) = \frac{1}{\sqrt{\det(2\pi\Sigma(\theta,\mathcal{T}))}} \exp\Big[-\frac{1}{2}\big(\vec{\mathbf{X}}-\vec{\mu}(\theta,\mathcal{T})\big)'\Sigma(\theta,\mathcal{T})^{-1}\big(\vec{\mathbf{X}}-\vec{\mu}(\theta,\mathcal{T})\big)\Big] \quad (1.7)$$

Seen as a function of θ, eq. 1.7 represents the likelihood of the given phylogenetic model evaluated at the trait data $\vec{\mathbf{X}}$. Phylogenetic models, which exhibit the above multivariate Gaussian form of the likelihood function, are called Gaussian phylogenetic models.

The efficient calculation of the model likelihood (eq. 1.7) is the main challenge for enabling the fast inference of Gaussian phylogenetic models given a big phylogenetic tree and trait data. The reason is that eq. 1.7 involves the construction and inversion of the $kN \times kN$ covariance matrix Σ. In Chapters 5 and 6, I'll show how these computationally heavy operations can be skipped for a specific sub-family of the Gaussian phylogenetic models.

1.4.4 *Inferring phylogenetic model parameters*

Inferring the parameters of a phylogenetic model of evolution consists in evaluating the space of model parameters θ with the goal to find a subset or a point, θ^* in this space that "fits best" to a given tree $\mathcal{T}$ and data $\mathbf{X}$ associated with its tips. This topic will be present in Chapters 3-7 of the thesis. I will study three types of model inference:

- maximum likelihood inference, consisting in maximizing the function$pdf(\vec{\mathbf{X}}|\theta,\mathcal{T})$ over θ. This type of inference will be used in Chapters 3, 4, and 7;

- Bayesian inference, consisting in finding a sample from the posterior distribution of θ given a the data, the tree and a prior distribution $P(\theta)$. This type of inference will be applied in Chapters 3 and 4. Since this type of inference is a time intensive task requiring millions of likelihood evaluations, in Chapter 5, I explore the possibility to speed-up the Bayesian inference of such a Gaussian phylogenetic model by parallelizing the likelihood evaluation.
- Maximum likelihood based model selection, which uses multiple maximum likelihood inferences for a set of "candidate" models in order selecting a best model. This type of inference will be the main subject in Chapter 7.

1.5 A NOTE ON THE REFERENCES

Since chapters 2-4 are included as original publications, the references in these chapters are at their ends (before Appendices). The references in the remaining chapters are at the end of the thesis.

Part II

PUBLICATIONS

2

INTRODUCTION TO PHYLOGENETICS

Published as

Joëlle Barido-Sottani, Veronika Bošková, Louis Du Plessis, Denise Kühnert, Carsten Magnus, **Venelin Mitov**, Nicola F. Müller, Jūlija PečErska, David A. Rasmussen Chi Zhang, Alexei J. Drummond Tracy A. Heath, Oliver G. Pybus, Timothy G. Vaughan and Tanja Stadler (2017).

Taming the BEAST — A Community Teaching Material Resource for BEAST 2 *Systematic Biology* 67(1):170-174, 2017

In 2016, the PhD students in the Computational Evolution (cEvo) group at ETH Zurich, with the support of senior post-doctoral students and professors from several other groups, organized a summer school in Bayesian Phylogenetics and Phylodynamics entitled "Taming the Beast" (inspired by the popular software for Bayesian phylogenetic inference BEAST 2). The original idea for this summer school came from Dr. Louis du Plessis, a former PhD student in the group. All of us took part in organizing the teaching materials for the lectures and seminars held during this summer school. This publication presents an online portal where all teaching materials from the summer school are publicly and freely available.

Following is the original publication in Systematic Biology from 2017.

Software for Systematics and Evolution

Syst. Biol. 67(1):170–174, 2018

DOI:10.1093/sysbio/syx060
Advance Access publication June 29, 2017

Taming the BEAST—A Community Teaching Material Resource for BEAST 2

JOËLLE BARIDO-SOTTANI[1,2,†], VERONIKA BOŠKOVÁ[1,2,†], LOUIS DU PLESSIS[1,3,†], DENISE KÜHNERT[1,2,4,†], CARSTEN MAGNUS[1,2,†], VENELIN MITOV[1,2,†], NICOLA F. MÜLLER[1,2,†], JŪLIJA PEČERSKA[1,2,†], DAVID A. RASMUSSEN[1,2,†], CHI ZHANG[1,2,†], ALEXEI J. DRUMMOND[5,‡], TRACY A. HEATH[6,‡], OLIVER G. PYBUS[3,‡], TIMOTHY G. VAUGHAN[5,‡], AND TANJA STADLER[1,2,*,§]

[1]*Department of Biosystems Science and Engineering, ETH Zürich, Mattenstrasse 26, 4058 Basel, Switzerland;* [2]*Swiss Institute of Bioinformatics (SIB), Quartier Sorge - Batiment Genopode, 1015 Lausanne, Switzerland;* [3]*Department of Zoology, University of Oxford, Peter Medawar Building South Parks Road Oxford, OX1 3SY, UK;* [4]*Department of Environmental Sciences, ETH Zürich, Universitätsstrasse 16, 8092 Zürich, Switzerland;* [5]*Centre for Computational Evolution, University of Auckland, New Zealand; and* [6]*Department of Ecology, Evolution, and Organismal Biology, Iowa State University, 2200 Osborn Dr., Ames, IA 50011 USA*

**Correspondence to be sent to: Department of Biosystems Science and Engineering, ETH Zürich, Mattenstrasse 26, 4058 Basel, Switzerland; E-mail: tanja.stadler@bsse.ethz.ch.*

[†]*These authors contributed equally.* [‡]*These authors contributed equally.*

[§]*Senior author.*

Received 21 December 2016; reviews returned 20 June 2017; accepted 25 June 2017
Associate Editor: David Bryant

Abstract.—Phylogenetics and phylodynamics are central topics in modern evolutionary biology. Phylogenetic methods reconstruct the evolutionary relationships among organisms, whereas phylodynamic approaches reveal the underlying diversification processes that lead to the observed relationships. These two fields have many practical applications in disciplines as diverse as epidemiology, developmental biology, palaeontology, ecology, and linguistics. The combination of increasingly large genetic data sets and increases in computing power is facilitating the development of more sophisticated phylogenetic and phylodynamic methods. Big data sets allow us to answer complex questions. However, since the required analyses are highly specific to the particular data set and question, a black-box method is not sufficient anymore. Instead, biologists are required to be actively involved with modeling decisions during data analysis. The modular design of the Bayesian phylogenetic software package BEAST 2 enables, and in fact enforces, this involvement. At the same time, the modular design enables computational biology groups to develop new methods at a rapid rate. A thorough understanding of the models and algorithms used by inference software is a critical prerequisite for successful hypothesis formulation and assessment. In particular, there is a need for more readily available resources aimed at helping interested scientists equip themselves with the skills to confidently use cutting-edge phylogenetic analysis software. These resources will also benefit researchers who do not have access to similar courses or training at their home institutions. Here, we introduce the "Taming the Beast" (https://taming-the-beast.github.io/) resource, which was developed as part of a workshop series bearing the same name, to facilitate the usage of the Bayesian phylogenetic software package BEAST 2. [Bayesian inference; MCMC; phylodynamics; phylogenetics.]

BEAST 2 IN A NUTSHELL

BEAST 2 (Bouckaert et al. 2014) is an open source cross-platform software package for analysing genetic sequences in a Bayesian phylogenetic framework. It occupies the same niche, and thus incorporates many of the same models, as other popular Bayesian evolutionary analyses platforms, including BEAST (Drummond and Rambaut 2007) (which we refer to here as BEAST 1 in order to distinguish it from BEAST 2), MrBayes (Huelsenbeck and Ronquist 2001), and RevBayes (Höhna et al. 2016). Although BEAST 2 is a complete redesign of the BEAST 1 software package, it retains a similar user interface and many core model components, including relaxed molecular clock models (Drummond et al. 2006), Bayesian skyline models for nonparametric coalescent analyses (Drummond et al. 2005; Heled and Drummond 2008), multispecies coalescent inference with *BEAST (Drummond and Heled 2010), and phylogeographical models (Lemey et al. 2009; 2010). Like in BEAST 1, an analysis is set up using input XML files. For most standard analyses, these files can be easily created using a graphical user interface (BEAUti 2).

The key difference in design philosophy between BEAST 1 and BEAST 2 is a greater emphasis in the latter on extensibility, resulting in a modular program built around a set of core components. This allows third-party developers to implement new methods as packages that can be added without rebuilding or redeploying BEAST 2. Through such packages, BEAST 2 provides a growing collection of new models not available in BEAST 1, such as flexible birth–death tree-priors (Stadler et al. 2013; Gavryushkina et al. 2014; Kühnert et al. 2016)

and structured coalescent models (Vaughan et al. 2014; De Maio et al. 2015), as well as updates to existing models, such as StarBEAST 2 (Ogilvie and Drummond 2016). A list of available models in BEAST 1 and BEAST 2 can be found at http://beast2.org/beast-features/. (Users should bear in mind that BEAST 2 is modular by design, and thus some third-party packages may not be listed.)

This modular design requires the BEAST 2 user to make active modeling choices, and it is no longer possible to simply perform a "default" analysis. This active involvement opens the door for analyses tailored specifically to particular data sets and questions, greatly increasing the power of the package. However, it also markedly increases the complexity and makes it easier to inadvertently introduce errors or use inappropriate models. This added complexity could also be daunting to novice users and may result in them preferring simpler, but less powerful, software packages. We will now briefly highlight the key steps required from the BEAST 2 user when running a data analysis.

At its core, BEAST 2 estimates rooted phylogenies ($\mathcal{T}$) from genetic sequencing data ($\mathcal{D}$), with branch lengths in units of calendar time (i.e., the phylogenies are time-trees). It concurrently estimates evolutionary parameters (θ), such as the substitution rate, and parameters describing population dynamics (η), such as speciation/extinction or transmission/recovery rates. For inference, BEAST 2 uses a Markov chain Monte Carlo (MCMC) algorithm to sample from the posterior distribution,

$$\Pr[\mathcal{T},\eta,\theta|\mathcal{D}] = \frac{\Pr[\mathcal{D}|\mathcal{T},\theta]\Pr[\mathcal{T}|\eta]\Pr[\eta]\Pr[\theta]}{\Pr[\mathcal{D}]}. \quad (1)$$

The output of an analysis is a log-file containing a sample of the states $(\mathcal{T},\eta,\theta)$ visited by the MCMC algorithm. After a so-called burn-in phase, each value $(\mathcal{T},\eta,\theta)$ is visited by the chain at a frequency proportional to its posterior probability, so the output of BEAST 2 (after eliminating the burn-in) is a set of samples from the posterior distribution. A recent book (Drummond and Bouckaert 2015) describes the general theory and design behind BEAST 2.

For the user to carry out a successful and correct analysis, several steps need to be performed carefully to analyze the data and answer the research question of interest. The researcher must specify a multileveled (i.e., hierarchical) model with several interacting components, including: (i) a suitable model describing the evolution of the sequence data on a time-tree, including the substitution and molecular-clock models ($\Pr[\mathcal{D}|\mathcal{T},\theta]$); (ii) a phylodynamic model describing the growth of the tree over time ($\Pr[\mathcal{T},\eta]$); and (iii) sensible prior distributions for each of the parameters of the evolutionary models ($\Pr[\theta]$ and $\Pr[\eta]$).

In addition to the model components, the researcher must also specify and fine-tune MCMC operators that propose new states for the model parameters $(\mathcal{T},\eta,\theta)$. By choosing appropriate proposal algorithms, an MCMC analysis is more likely to sample the posterior distribution efficiently. Finally, once the MCMC chain has sampled a sufficient number of states, the researcher must assess whether the chain has converged and recovered a meaningful signal from the data.

Consequently, the user is challenged with a myriad of choices on the road to a successful analysis. Although many potential pitfalls exist, a simple but solid understanding of the theory behind Bayesian phylogenetic inference can help guide new users through an analysis to reach sound conclusions.

"Taming the Beast" for the User Community

In June 2016, we organized a "Taming the BEAST" workshop in Engelberg, Switzerland, aimed at fostering interaction between BEAST 2 users and developers. The workshop was organized by graduate students and postdoctoral researchers in the Computational Evolution group at ETH Zürich (https://www.bsse.ethz.ch/cevo, with generous financial support from ETH Zürich) and was a mix of lectures by invited speakers (A.J.D., T.A.H., O.G.P., T.G.V., and T.S. were invited speakers.) and hands-on tutorials run by the organisers. (J.B.-S., V.B., L.d.P., D.K., C.M., V.M., N.F.M., J.P., D.A.R., and C.Z. organized the tutorial sessions.) Participants had the opportunity to learn how to use BEAST 2 with help from the developers and to discuss questions specific to their research with other experienced scientists. For the developers, such a workshop provides direct feedback from users on ease-of-use, identifying specific issues and discovering the needs and wishes of the community for future software and methods development.

The workshop was met with great enthusiasm from researchers already using or planning to use BEAST 2, ranging from students to established PIs. (Although originally envisioned for graduate students only, many postdoctoral researchers, some lecturers, and a few professors applied for the workshop as well. Due to the limited capacity and resources, out of 75 applications, we selected 36 participants from 14 countries and 28 universities.) The positive feedback from the participants (see Fig. 1), the overwhelming support from the community and the demand for further workshops has provided motivation to initiate a series of "Taming the BEAST" workshops. At the time of writing, a second successful edition of "Taming the Beast" was run on Waiheke island (New Zealand) in February 2017 and a third edition will take place in July 2017 in London. Further editions are planned for 2018 in Switzerland, and for 2019 and 2020 in locations that are yet to be determined. (We secured funding from ETH Zürich to support the workshop series in 2017–2020.) Each workshop is intended as a global event, allowing users and developers from around the world to meet and share knowledge.

To ensure these resources are available to the community, we have set up a website (https://taming-

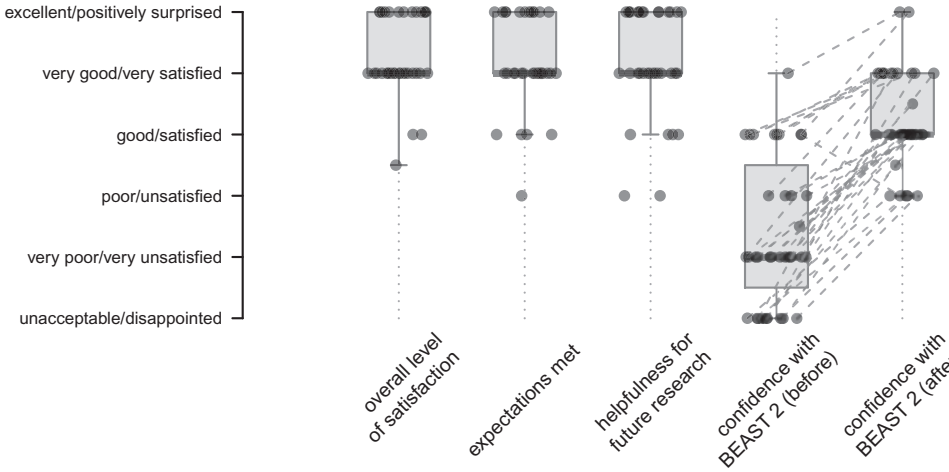


FIGURE 1. Boxplot showing the feedback received from 35 respondents (out of 36 workshop participants) on 5 feedback questions. Of the 35 respondents, all but 3 indicated that they would definitely recommend the workshop to a colleague.

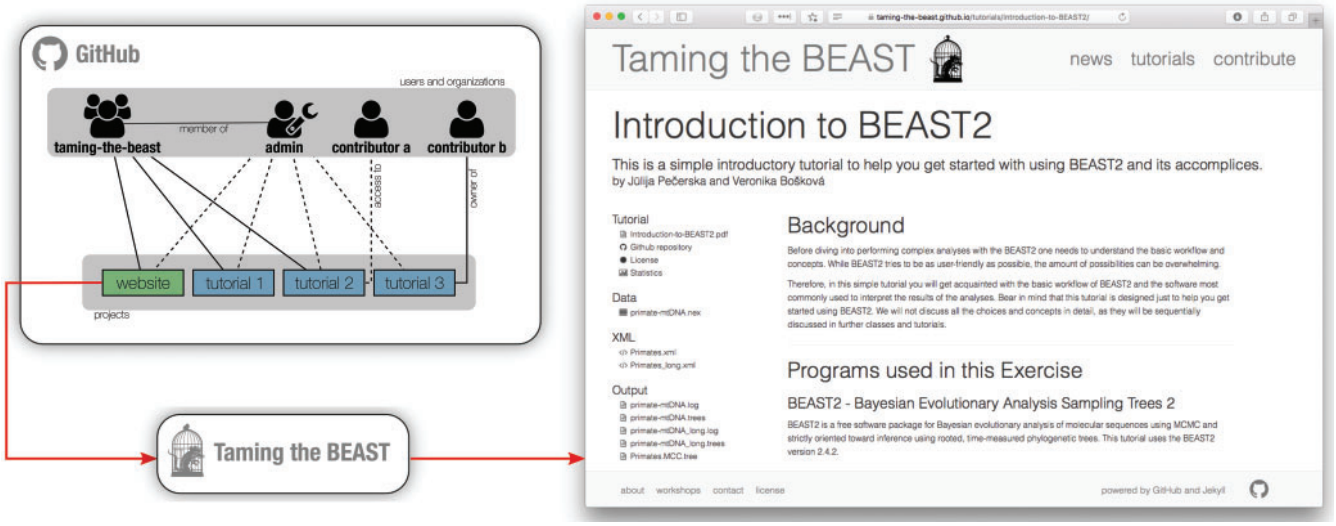


FIGURE 2. Structure of the Taming the BEAST web resource as hosted on GitHub. The diagram on the left shows three possibilities for tutorials available on the website. On the diagram solid lines indicate ownership and dashed lines access. Tutorial 1 is owned by the taming-the-beast organization on GitHub, and does not have any external contributors. Tutorial 2 was created by contributor a, but ownership has been transferred to taming-the-beast. Tutorial 3 was created by contributor b, who has retained ownership. In all three cases, it is essential that at least one of the website administrators has access to the tutorial. The website itself is also hosted on GitHub as a project. When a user visits the website tutorials appear as on the right of the figure. The left panel contains links to a printable PDF version of the tutorial, the data file (or files) used in the tutorial, example BEAST 2 XML files, examples output files and a link to the GitHub repository of the tutorial. Recent changes to the tutorial are also listed.

the-beast.github.io/) with the same name as the workshop series to serve as a platform for collating a comprehensive and cohesive set of BEAST 2 tutorials (see Fig. 2). By providing a set of well-curated tutorials, "Taming the BEAST" offers researchers the resources necessary to learn how to perform analyses in BEAST 2. In addition to tutorials provided by the BEAST 2 developers, this resource page also contains all of the materials (lecture slides, tutorials, data, and example outputs) used during the first two "Taming the BEAST" workshops in Switzerland and New Zealand. These materials will be updated and extended for future editions of the workshop. Tutorials are released under a license that gives anyone the right to freely use (and modify) tutorials for courses or workshops, as long as appropriate credit is given and the updated material is licensed in the same fashion. (By default we use a Creative Commons Attribution 4.0 license, however the exact license to be used is determined by the tutorial's authors.) We hope that these open resources will encourage other research groups/universities to host and organize their own "Taming the BEAST" workshops. As a community resource, the "Taming the BEAST" website will maintain a list of workshops, and tutorial developers are available to provide support to organizers.

Contributing to Taming the Beast

In keeping with the BEAST 2 design philosophy, we designed the website to have a modular, extensible architecture. Each tutorial is stored in its own GitHub (http://www.github.com) repository, where it is bundled with all of the supporting data and scripts needed to run the tutorial, as well as example output files. This makes it possible for anyone with a GitHub account to raise issues and suggest edits or extensions to tutorials. Similarly, it is also possible for external contributors to submit new tutorials to the website. We provide a template tutorial and comprehensive documentation to help potential contributors get started.

By providing a "Taming the Beast" platform that allows issues to be raised and content to be edited, we hope that the community will play an active role in curating tutorials. We further envision these resources will continue to grow as the community contributes more tutorials. For instance, the developers of a new BEAST 2 package will be able to add a tutorial for their package to the "Taming the BEAST" site, where it will be accessible in a central location, along with other BEAST 2 tutorials, making it easier for users to become familiar with their package.

Because tutorials are stored in GitHub repositories that track change history, all contributors can receive proper credit for their work. Furthermore, authors of new tutorials can retain ownership of their tutorials after publication. In addition, GitHub tracks traffic to tutorials over time and makes it easy for users to interact with authors, giving authors a measure of their work's impact within the community. Finally, because of the distributed nature of the website, it is robust to changes in any single repository, making it easy to update or add individual tutorials.

Summary

The tutorials on the "Taming the Beast" website allow users to learn about the entire BEAST 2 analysis pipeline, with most tutorials focusing on a particular model component or a single BEAST 2 package. The website provides immediate access to the materials that guide users in the application of a range of models to their own data. In addition, there are tutorials on postprocessing, interpreting results, as well as troubleshooting. We will ensure the maintenance of the website and incorporation of new tutorials through two to three responsible people from the Computational Evolution group at ETH Zürich as well as collaborating groups acting as website administrators. The administrators of the website can be reached via tamingthebeast@bsse.ethz.ch.

We hope that the "Taming the BEAST" platform will allow new BEAST 2 users to accelerate their learning process and to successfully "tame" the BEAST. At the same time, we hope that it will serve as a central repository of teaching materials that will allow BEAST 2 developers and users to exchange knowledge about how to effectively teach the use of BEAST 2. Finally, this platform will hopefully further encourage developers to share their own materials with the wider community.

Acknowledgments

First and foremost we would like to express our immense gratitude to the community for the overwhelmingly positive response both before the first workshop (in the form of letters of support and interest) and after the workshop (in helping us turn it into a series of recurring workshops). We would also like to thank the BEAST 2 core developers for supporting our initiatives and helping us to run the workshop smoothly, in particular Walter Xie and Remco Bouckaert who tested tutorials and implemented last minute bug-fixes. We further acknowledge generous support from ETH Zürich through the Swiss University Conference (SUK) program. The website architecture is based on Trevor Bedford's lab website. Many thanks to Trevor for making his code publicly available! O.G.P. wishes to thank Andrew Rambaut for his contributions to lecture slides. Further, we would like to thank the speakers of the second workshop, Simon Ho, David Bryant, Remco Bouckaert, Huw Ogilvie, and David Duchêne, as well as Carmella Lee for organizing the logistics of the second workshop. Finally, we would like to thank David Bryant and an anonymous reviewer for valuable comments on the article.

Author's Contributions

J.B.-S., V.B., L.d.P., V.M., and J.P. wrote and submitted the SUK application for starting the "Taming the BEAST" workshop series, with substantial support of C.M. and D.A.R. The first workshop was organized by the whole Computational Evolution group (led by J.B.-S., V.B., and L.d.P.). J.B.-S., V.B., L.d.P., D.K., C.M., V.M., N.F.M., J.P., D.A.R., C.Z., A.J.D., T.A.H., O.G.P., T.G.V., and T.S. wrote the tutorials and/or lecture slides for teaching. L.d.P. created the figures, set up the web resource and GitHub repositories and is the corresponding person regarding these online resources. L.d.P., J.B.-S., V.B., and T.S. wrote the article.

References

Bouckaert R., Heled J., Kühnert D., Vaughan T., Wu C.-H., Xie D., Suchard M.A., Rambaut A., Drummond A.J. 2014. BEAST 2: a software platform for Bayesian evolutionary analysis. PLoS Comput. Biol. 10(4):e1003537.

De Maio N., Wu C.-H., O'Reilly K.M., Wilson D. 2015. New routes to phylogeography: a Bayesian structured coalescent approximation. PLoS Genet. 11(8):e1005421.

Drummond A.J., Bouckaert R.R. 2015. Bayesian evolutionary analysis with BEAST. Cambridge, UK: Cambridge University Press.

Drummond A.J., Heled J. 2010. Bayesian inference of species trees from multilocus data. Mol. Biol. Evol. 27(3):570–580.

Drummond A.J., Ho S.Y.W., Phillips M.J., Rambaut A. 2006. Relaxed phylogenetics and dating with confidence. PLOS Biol. 4(5):e88.

Drummond A.J., Rambaut A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol. Biol. 7(1):1.

Drummond A.J., Rambaut A., Shapiro B., Pybus, O.G. 2005. Bayesian coalescent inference of past population dynamics from molecular sequences. Mol. Biol. Evol. 22(5):1185–1192.

Gavryushkina A., Welch D., Stadler T., Drummond A.J. 2014. Bayesian inference of sampled ancestor trees for epidemiology and fossil calibration. PLoS Comput. Biol. 10(12):e1003919.

Heled J., Drummond A.J. 2008. Bayesian inference of population size history from multiple loci. BMC Evol. Biol. 8:289.

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

Huelsenbeck J.P., Ronquist F. 2001. Mrbayes: Bayesian inference of phylogenetic trees. Bioinformatics 17(8):754–755.

Kühnert D., Stadler T., Vaughan T.G., Drummond A.J. 2016. Phylodynamics with migration: a computational framework to quantify population structure from genomic data. Mol. Biol. Evol. 33(8):2102–2116.

Lemey P., Rambaut A., Drummond A.J., Suchard M.A. 2009. Bayesian phylogeography finds its roots. PLOS Comput. Biol. 5(9):e1000520.

Lemey P., Rambaut A., Welch J.J., Suchard M.A. 2010. Phylogeography takes a relaxed random walk in continuous space and time. Mol. Biol. Evol. 27(8):1877–1885.

Ogilvie, H.A., Bouckaert, R.R., Drummond, A.J. 2017. StarBEAST2 brings faster species tree inference and accurate estimates of substitution rates. Mol. Biol. Evol. doi: 10.1093/molbev/msx126. [Epub ahead of print].

Stadler T., Kühnert D., Bonhoeffer S., Drummond A.J. 2013. Birth–death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis c virus (HCV). Proc. Natl. Acad. Sci. USA 110(1): 228–233.

Vaughan T.G., Kühnert D., Popinga A., Welch D., Drummond A.J. 2014. Efficient Bayesian inference under the structured coalescent. Bioinformatics 30(16):2272–9.

3

ESTIMATING THE HERITABILITY OF PATHOGEN TRAITS

Published as

Venelin Mitov and Tanja Stadler (2018). A Practical Guide to Estimating the Heritability of Pathogen Traits. *Molecular Biology and Evolution* 6:9.

Since the work of Alizon et al. (2010), proposing a PCM-based approach to estimating the heritability of virulence of HIV infections, several laboratories have developed similar techniques and applied them to different cohorts of HIV patients. With an order of magnitude difference between the lowest and the highest estimates, there was a controversy in the field. Some authors supported the hypothesis of zero or negligible heritability, meaning that the virus strain infecting a patient does not have an effect on the time it would take for the patient to develop AIDS, in the absence of therapy. Other authors believed that the virus was playing a dominant role or, at least, the virulence resulted from the interplay between the virus and the immune system specific for every infection. A strong statistical support for the first hypothesis was shown in Hodcroft et al. (2014), who estimated HIV set point viral load (spVL) heritability in the UK subtype B cohort (N=8468) to less than 6%. In this article, I applied different heritability estimators on the same dataset and to a number of epidemic simulations following a mechanistic model of within- and between-host viral dynamics during an epidemic. Based on these results, I state that the estimate of 6% is a negatively biased estimator of the spVL-heritability, due to the fact that the assumed Brownian motion model for the viral evolution did not fit to the pattern exhibited by the comparative data.

Following is the original publication, which appeared in Molecular Biology and Evolution in the early 2018.

A Practical Guide to Estimating the Heritability of Pathogen Traits

Venelin Mitov*[,1,2] and Tanja Stadler[1,2]

[1]Department of Biosystems Science and Engineering, ETH Zürich, Basel, Switzerland

[2]Swiss Institute of Bioinformatics, Switzerland

***Corresponding author:** E-mail: vmitov@gmail.com.

Associate editor: Sergei Kosakovsky Pond

Abstract

Pathogen traits, such as the virulence of an infection, can vary significantly between patients. A major challenge is to measure the extent to which genetic differences between infecting strains explain the observed variation of the trait. This is quantified by the trait's broad-sense heritability, H^2. A recent discrepancy between estimates of the heritability of HIV-virulence has opened a debate on the estimators' accuracy. Here, we show that the discrepancy originates from model limitations and important lifecycle differences between sexually reproducing organisms and transmittable pathogens. In particular, current quantitative genetics methods, such as donor–recipient regression of surveyed serodiscordant couples and the phylogenetic mixed model (PMM), are prone to underestimate H^2, because they neglect or do not fit to the loss of resemblance between transmission partners caused by within-host evolution. In a phylogenetic analysis of 8,483 HIV patients from the United Kingdom, we show that the phenotypic correlation between transmission partners decays with the amount of within-host evolution of the virus. We reproduce this pattern in toy-model simulations and show that a phylogenetic Ornstein–Uhlenbeck model (POUMM) outperforms the PMM in capturing this correlation pattern and in quantifying H^2. In particular, we show that POUMM outperforms PMM even in simulations without selection—as it captures the mentioned correlation pattern—which has not been appreciated until now. By cross-validating the POUMM estimates with ANOVA on closest phylogenetic pairs, we obtain $H^2 \approx 0.2$, meaning $\sim$20% of the variation in HIV-virulence is explained by the virus genome both for European and African data.

Key words: **HIV, set-point viral load (spVL), donor–recipient regression, ANOVA, phylogenetic mixed model, Ornstein–Uhlenbeck.**

Article

Introduction

Pathogens transmitted between donor and recipient hosts are genetically related much like children are related to their parents through inherited genes. This analogy between transmission and biological reproduction has inspired the use of heritability (H^2)—a term borrowed from quantitative genetics (Falconer and Mackay 1996; Lynch and Walsh 1998; Hartl and Clark 2007) to measure the contribution of pathogen genetic factors to pathogen traits, such as virulence, transmissibility, and drug-resistance of infections.

Two families of methods have been used to estimate the heritability of a pathogen trait in the absence of knowledge about its genetic basis:

- Resemblance estimators measuring the relative trait-similarity within groups of transmission-related patients. Common methods of that kind are linear regression of donor–recipient (DR) couples (Fraser et al. 2014; Leventhal and Bonhoeffer 2016) and analysis of variance (ANOVA) of patients linked by (near-)identity of carried strains (Anderson et al. 2010; Shirreff et al. 2013).
- Phylogenetic comparative methods measuring the so called phylogenetic heritability, that is, the association between observed trait values from patients and their (approximate) transmission tree inferred from pathogen sequences. Common examples of such methods are the Felsenstein's independent contrasts (Felsenstein 1985), the phylogenetic mixed model (PMM) (Housworth et al. 2004), and the Pagel's λ (Freckleton et al. 2002).

Most of these methods have been applied in studies of the viral contribution to virulence in an HIV infection (Tang et al. 2004; Alizon et al. 2010; Hecht et al. 2010; Hollingsworth et al. 2010; van der Kuyl et al. 2010; Lingappa et al. 2013; Shirreff et al. 2013; Yue et al. 2013; Fraser et al. 2014; Hodcroft et al. 2014; Bonhoeffer et al. 2015; Leventhal and Bonhoeffer 2016; Bachmann et al. 2017; Bertels et al. 2018; Blanquart et al. 2017). To quantify the virulence of an HIV infection, the above studies have used measurements of the $\log_{10}$ set point viral load, lg(spVL)—the amount of virions per blood-volume stabilizing in HIV patients at the beginning of the asymptomatic phase and best-predicting its duration (Mellors et al. 1996). In the view of discrepant reports of lg(spVL)-heritability, many authors have questioned the accuracy of the existing methods and have proposed various adaptations of these methods in order to overcome potential pitfalls, such as false model assumptions (e.g., neutral evolution and ultrametricity

of transmission trees) and imperfections in the data (e.g., small data size, presence of cofactors, and measurement error) (Shirreff et al. 2013; Fraser et al. 2014; Hodcroft et al. 2014; Leventhal and Bonhoeffer 2016; Mitov and Stadler 2016; Bachmann et al. 2017; Bertels et al. 2018; Blanquart et al. 2017). Despite these efforts, to date, there is no consensus about the root cause of the discrepancy in lg(spVL)-heritability estimates and there is little reuse of the tools previously implemented, making it hard to compare the estimates from different studies.

In the remainder of the introduction, we consider the definition of broad-sense heritability from the point of view of the key differences between sexually reproducing organisms and clonally transmitted pathogens. Then, in New Approaches, we introduce a model of an epidemic that allows exploring how one of these differences—the within-host evolution of pathogens—affects most of the currently used estimators of heritability. In Results, we compare these estimators based on simulations of the above model and report an analysis of spVL data from a large HIV cohort. In the light of these results, we designate the most reliable estimators of pathogen trait heritability and establish a lower bound for the viral genetic contribution to set-point viral load.

Differences between Pathogens and Sexual Species When Estimating Heritability

According to quantitative genetics theory, the *broad-sense heritability*, H^2, of a quantitative trait is defined in the context of a population of organisms as the ratio of the genotypic over phenotypic variance:

$$H^2 = \mathrm{Var}(G)/\mathrm{Var}(z), \quad (1)$$

where z denotes the phenotypic value and G denotes the genotypic value assigned to each individual in the population (Falconer and Mackay 1996; Lynch and Walsh 1998; Hartl and Clark 2007). In the case of epidemics, the population represents a sample of hosts, that is, organisms infected by a given type of pathogen. The phenotypic value, z, represents a numerical trait resulting from the infection, and the genotypic value, G, is defined for each pathogen genotype (strain), as the phenotypic value to be expected if it would be measured in a randomly chosen host infected with this strain.

In a large enough population with fully known pathogen genotypes, H^2 could be measured by the direct heritability estimator—the coefficient of determination, R^2_{adj}, obtained over a grouping of the population by genotype. In practice though, this is impossible, because population sizes are small compared with large numbers of (usually unknown) genotypes. To tackle this problem, pathogenecists have relied on the apparent analogy between parent–offspring couples in sexually reproducing populations and DR couples in infected populations. This analogy has motivated the use of correlation measures, such as the DR regression slope, b, and the intraclass correlation in phylogenetic pairs, r_A, to estimate the heritability of pathogen traits (Anderson et al. 2010; Shirreff et al. 2013; Fraser et al. 2014; Leventhal and Bonhoeffer 2016). However, three differences between the lifecycles of clonally transmitted pathogens and sexually reproducing organisms challenge this approach:

Asexual Haploid Nature of Pathogen Transmission

The first difference is that, unlike the reproduction of diploid organisms, the transmission of a pathogen from a donor to a recipient is more similar to asexual (haploid) reproduction, because, typically, whole pathogens get transferred between hosts.

Partial Quasispecies Transmission

The second difference is that the transmitted proportion of genetic information characterizing the pathogen in the donor is unknown and varying between transmission events. For example, for slowly evolving bacteria such as *Mycobacterium tubercolosis* (Mtb), transmission can be clonal (Bjorn-Mortensen et al. 2016), whereas, for rapidly evolving retroviruses like HIV, transmission is often accompanied by bottlenecks causing only a tiny sample of the large and genetically diverse virus population in the donor (a.k.a., quasispecies) to penetrate and survive in the recipient (Keele et al. 2008).

Within-Host Pathogen Evolution

The third difference involves the change in phenotypic value due to within-host pathogen mutation and recombination. Although genetic change is rare during the lifetime of animals and plants and its phenotypic effects are typically delayed to the offspring generations, it constitutes a hallmark in the lifecycle of pathogens and causes a gradual or immediate phenotypic change such as increasing virulence, immune escape, or drug resistance (fig. 1).

The net outcome of these differences is that unlike family members, for which the amount of genetic overlap is a known constant, for example, 50% for a parent–child couple, the genetic overlap between the two quasispecies in a DR couple is an unknown variable. If there were full quasispecies transmission and no within-host evolution, the pathogen populations found in a donor and a recipient at any moment after a transmission event would be similar to identical twins raised in separate environments. By analogy with twins, any measure of the trait correlation in transmission couples, such as b and r_A, would estimate the broad-sense heritability, H^2 (Lynch and Walsh 1998). However, the partial quasispecies transmission and the within-host evolution taking place in the time between transmission and measurement can lead to a change in the correlation between couple members without affecting H^2 at the population level. We presume that this issue has been at the origin of the discrepancy in previous reports of lg(spVL)-heritability. In particular, the applied methods vary substantially in how they account for the within-host evolution taking place between transmission and measurement: some of them neglect it (Shirreff et al. 2013; Leventhal and Bonhoeffer 2016); others diminish its effect through preferential sampling of patients in the early phase of infection

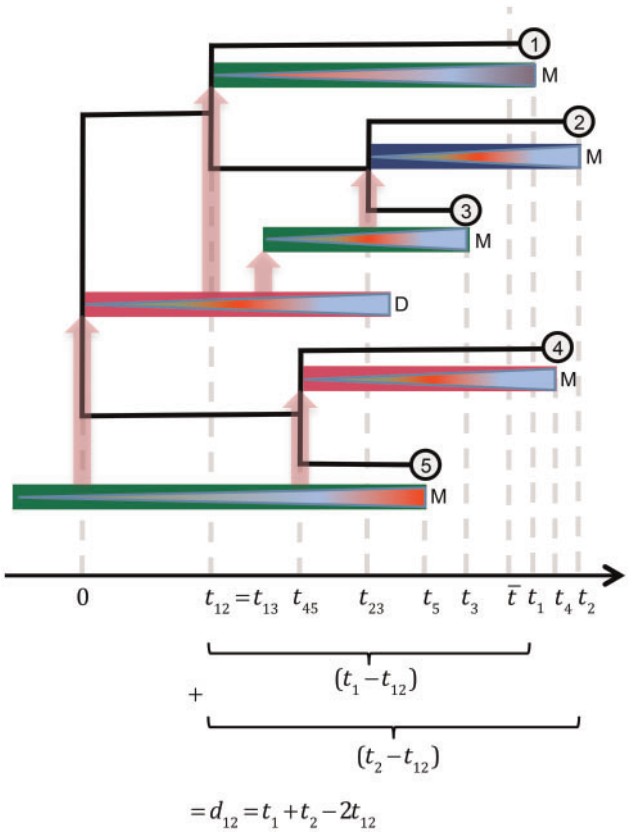


Fig. 1. A schematic representation of an epidemic. Colored rectangles represent infectious periods of hosts, different colors corresponding to different host types. Triangles inside hosts represent pathogen quasispecies, change of color indicating substitution of dominant strains. Capital letters denote host-events: M: diagnosis followed by immediate phenotype measurement, treatment and quarantine for the host; D: host death. The transmission tree connecting the measured hosts is drawn in black. Notice that, due to incomplete sampling, there is no one-to-one correspondance between transmission events and branching points on the tree. By convention, the time origin is at the root of the tree and the time is assumed to increase toward the the tips. We denote by t_i the time distance from the root to tip i. The mean root-tip distance is denoted by $\bar{t}$. For each couple of tips, i and j, we denote by t_{ij} the time distance from the root to their most recent common ancestor (mrca) and by d_{ij} their phylogenetic distance. For clarity, we show how d_{ij} can be expressed in terms of t_{ij} and the root-tip times, t_i and t_j. Couples of tips that are each other's closest tip by phylogenetic distance, for example, (2, 3) and (4, 5), are called "phylogenetic pairs" (PPs). In balanced trees, PPs tend to coincide with pairs of tips descending from the same parent node (a.k.a., siblings or "cherries").

or transmission couples shortly after seroconversion (Hecht et al. 2010; Hollingsworth et al. 2010); third ones attempt to account for it by taking advantage of stochastic models of trait evolution, such as Brownian motion (BM) (Alizon et al. 2010; Hodcroft et al. 2014) or Ornstein–Uhlenbeck (OU) (Mitov and Stadler 2016; Bertels et al. 2018; Blanquart et al. 2017). In the next section, we introduce a simulation based method allowing for within-host evolution, which enables comparing these methods against the direct heritability estimator, R^2_{adj}.

New Approaches

A Toy-Model of an Epidemic

We propose a simulation based method for evaluating different heritability estimators. Our approach differs substantially from previous simulation studies, where the pathogen genotype is equivalent to the genotypic value, G, and is modeled by a continuous branching stochastic process evolving along a given transmission tree (Alizon et al. 2010; Shirreff et al. 2013; Hodcroft et al. 2014; Leventhal and Bonhoeffer 2016). In contrast, we implement a more explicit model in which the pathogen genotype represents a randomly mutating sequence of gene variants (alleles) and the trait value results from the interaction between the pathogen genotype and the host. The main advantages of this approach are 1) the possibility to compare different estimates of H^2 to its true value obtained from the direct estimator, R^2_{adj}, and 2) the possibility to study the effect of within-host mutation and measurement delay on all estimates. As a limitation, the proposed model omits coexistence of strains within a host and partial quasispecies transmission, because of their complexity and the current lack of empirical knowledge and data (see Discussion). For this reason and because of its minimalistic design, we refer to this model as a "toy-model."

In the toy-model, we think of an infection as an asexually reproducing haploid organism. The environment for this organism is the infected host, and the reproduction represents the clonal transmission of the infecting strain to other susceptible hosts. The pathogen has a genome composed of a finite number of loci, which mutate sporadically during infection, resulting in mutant strains. Depending on the within-host fitness of a mutant, it can be eliminated or it can immediately substitute the strain currently invading the host. A trait, z, is determined by the additive effects and epistatic interactions between the alleles at the loci in the genome as well as the interaction between these alleles and the host immune system. The immune system represents a combination of an immutable host type interacting in a predefined way with each possible strain and a randomly drawn host-specific effect, summarizing the unknown effects of other host-related factors, such as age, sex, and habitat. We assume two equally frequent host-types and two trait-determining loci in the pathogen genotype with $M_1=3$ and $M_2=2$ possible alleles at each locus. Thus, there are six possible strains and a total of 12 host type×strain combinations (fig. 2A).

The dynamics of the model combine within-host events, such as strain mutation and substitution, and between-host events, such as transmission, natural, and pathogen-induced death as well as diagnosis followed by immediate uninfectiousness, recovery, and immunity for the patient. These events are modeled as Poisson processes for every infected individual (fig. 2B). The between-host dynamics are inspired from a classical Susceptible-Infected-Recovered (SIR) model with finite population size (ch. 1 in Keeling and Rohani 2007). The main difference with this epidemiological model is that the rate of transmission and the expected infectious period for an infected host can depend on the current trait value and

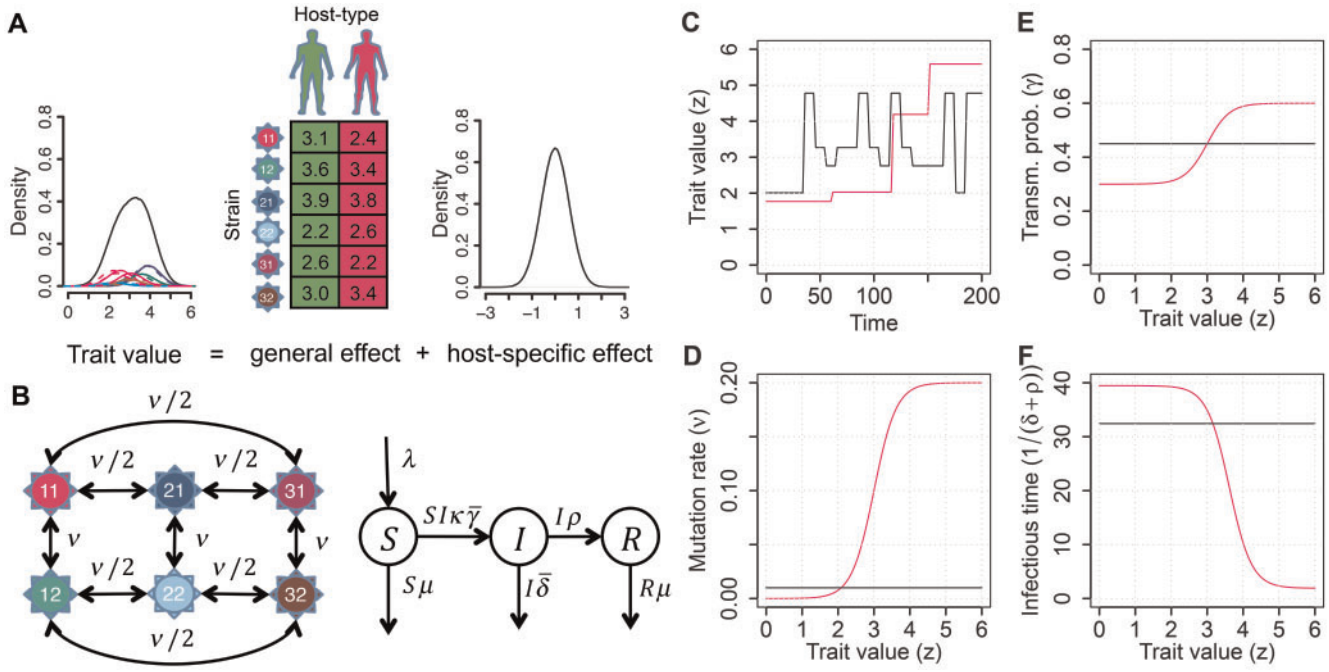


Fig. 2. A toy model of an epidemic with within-host mutation and SIR dynamics. (*A*) A pathogen trait represents the sum of a general <host type×strain> effect and a normally distributed host-specific effect. Pathogen strains are denoted by the alleles at the two loci, for example, "31" stays for allele 3 at locus 1 and allele 1 at locus 2. The density of the trait in a population of hosts represents a mixture of normal densities corresponding to the host type×strain combinations scaled by their relative frequencies. (*B*) Within a host (left), each locus of the infecting strain mutates at a rate ν; horizontal or curved arrows denote mutations at the first locus, vertical arrows denote mutations at the second locus; the rates above the arrows correspond to the per locus mutation rate (ν) divided by the number of possible other alleles at the locus. At the between-host level (right), the alive population is divided into a Susceptible, Infected, and Recovered compartments, letters *S*, *I* and *R* denoting the corresponding proportions in the population at a given time. New individuals become susceptible at a constant rate λ; risky contacts occur at rate $SI\kappa$, where κ denotes the individual contact rate; a risky contact can result in a new infection with probability γ, $\bar{\gamma}$ denoting the mean of the transmission probabilities of all infected hosts at a given time; a host is removed from the infected compartment in the events of death (occuring at rate δ) or diagnosis (occurring at rate ρ); diagnosis is followed by immediate treatment, recovery, and lifelong immunity for the patient; healthy hosts leave the *S* and *R* compartments at a constant rate μ. (*C*) An example time-course of the trait value within a host—the value changes instantaneously with strain mutation; in the "neutral" case (black), the trait can change upward or downward; in the "select" case (magenta), only positive changes are possible (mutants resulting in a lower trait value can not substitute the current strain). (*D*–*F*) The per locus mutation rate (ν), the per risky contact transmission probability (γ) and the expected infectious time ($1/(\delta+\rho)$) are defined as constants in the "neutral" case (black) or as functions of the trait value in the "select" case (magenta) (supplementary table S2, Supplementary Material online).

are subject to change with a substitution of the dominant strain within the host (magenta curves on fig. 2*D*–*F*). For each class of events (within- and between-host), we define two modes:

- neutral: events occur at rates defined as global constants mimicking neutrality (i.e., lack of selection) with respect to *z* (black lines on fig. 2*C*–*E*). For within-host events, it is assumed that a mutation of the pathogen is followed by instantaneous substitution of the mutant for the current dominant strain, regardless of the induced change in *z* (black line on fig. 2*C*);
- select: within hosts, it is assumed that a mutation of the pathogen is followed by instantaneous substitution only if it results in a higher *z* (magenta line on fig. 2*C*). Borrowing the approach from (Fraser et al. 2007), the rates of transmission and within-host mutation are defined as increasing Hill functions of 10^z, whereas the infectious time period is defined as a decreasing Hill function of 10^z, thus mimicking increasing per capita transmission- and pathogen-induced mortality for higher *z* (magenta lines on fig. 2*D*–*F*).

By combining different modes of dynamics at the within- and between-host levels the model can reproduce some popular hypotheses of pathogen evolution. For example, the combination of select within-host mode with select between-host mode simulates selection for optimal transmission potential (Fraser et al. 2007; Stearns and Koella 2007). This allows to evaluate the combined effect of selection and within-host trait evolution on various estimators of heritability.

Results

In this section, we use empirical data and simulations of the toy-model to show that most of the heritability estimators borrowed from classical quantitative genetics are prone to significant bias, because they neglect or inaccurately model the change in resemblance between transmission partners

caused by within-host evolution of the pathogen. Based on the toy-model simulations, we designate the intraclass correlation in the closest phylogenetic pairs (CPPs) and the phylogenetic heritability, $H^2_{\mathrm{OU}}(\bar{t})$, measured by the phylogenetic Ornstein–Uhlenbeck mixed model (POUMM) (Mitov and Stadler 2016; Blanquart et al. 2017) as the most reliable estimators of pathogen trait heritability. Based on applying these estimators to a large HIV cohort, we establish a lower bound for the lg(spVL)-heritability.

Through the rest of the article, we use the symbol d_{ij} to denote the phylogenetic distance between two tips, i and j, on a transmission tree (fig. 1). d_{ij} summarizes the total evolutionary distance between two infected hosts at the moment of measuring the trait value and is measured in substitutions per site for real trees and arbitrary time units for simulated trees. We begin our report with a result from HIV data demonstrating the relevance of within-host evolution for estimating heritability.

The lg(spVL) Correlation in HIV Phylogenetic Pairs Decreases with d_{ij}

We used one-way analysis of variance (ANOVA, r_{A}) and Spearman correlation (r_{Sp}) to estimate the correlation in phylogenetic pairs (PP) extracted from a recently published transmission tree of 8,483 HIV patients (Hodcroft et al. 2014). As defined in Shirreff et al. (2013), phylogenetic pairs represent pairs of tips in the transmission tree that are mutually nearest to each other by phylogenetic distance (d_{ij}) (fig. 1). We ordered the PPs by d_{ij} and split them into ten strata of equal size (deciles), evaluating the correlation between pair trait values (r_{A} and r_{Sp}) in each stratum. The point estimates and the 95% confidence intervals (CI) are shown with black and magenta points and error bars on figure 3. Dashed horizontal bars denote the 95% CI for r_{A} evaluated on all phylogenetic pairs. Despite some irregularities, there is a well pronounced pattern of decay in the correlation—strata to the left (small d_{ij}) tend to have higher r_{A} values than strata to the right (big d_{ij}). The values of r_{A} closely matched the values from other correlation estimators, such as DR (b) and the Pearson product mean correlation (r) (results not shown). We performed ordinary least squares regressions (OLS) of the values r_{A,D_k} and r_{Sp,D_k} on the mean phylogenetic distance, $\bar{d}_{ij,k}$, in each stratum, $k = 1, \ldots, 10$. The slopes of both regressions were significantly negative ($P<0.05$) and are shown as black and magenta lines on figure 3. Similar slopes were obtained when using other stratifications of the data (supplementary fig. S1, Supplementary Material online).

The above result shows that the value of a heritability estimator based on the correlation within phylogenetic pairs (including DR couples) depends strongly on d_{ij}. Another issue of all estimators of H^2 using the correlation in phylogenetic or DR pairs is that the underlying statistical methods require independence between the pairs—the trait values in one pair should not influence or be correlated with the trait values in any other pair. This assumption is not valid in general, due to the phylogenetic relationship between all patients. One way to mitigate the effects of phylogenetic relationship between pairs is to limit the analysis to the closest pairs (i.e., pairs, for which d_{ij} does not exceed some user specified threshold). This approach has the drawback of omitting much of the data from the analysis. As an alternative taking advantage of the entire tree, it is possible to correct for the phylogenetic relationship by using a phylogenetic comparative method (PCM). PCMs attempt to solve both of the above problems, because they 1) incorporate the branch lengths in the transmission tree to model the variance–covariance structure of the data and 2) correct for the phylogenetic correlation when estimating evolutionary parameters or the phylogenetic heritability of the trait (Felsenstein 1985; Housworth et al. 2004; Alizon et al. 2010). These advantages of the PCMs come at the price of assuming a specific stochastic process as a model of the trait evolution along the tree. In the next subsection, we show that assuming an inappropriate process for the trait evolution can cause a significant bias in the estimate of phylogenetic heritability.

A Brownian Motion Process Cannot Reproduce the Decay of Correlation in the UK Data

We implemented a maximum likelihood and a Bayesian fit of the PMM (Lynch 1991; Housworth et al. 2004) and its extension to an Ornstein–Uhlenbeck model of evolution (POUMM) (Hansen 1997; Mitov and Stadler 2016; Blanquart et al. 2017). The PMM and the POUMM assume an additive model of the trait values, $z(t) = g(t) + e$, in which $z(t)$ represents the trait value at time t for a given lineage of the tree, $g(t)$ represents a heritable (genotypic) value at time t for this lineage and e represents a nonheritable contribution summarizing the effects of the host and his/her environment on the trait and the measurement error. The only difference between the two models is their assumption about the evolution of $g(t)$ along the branches of the tree—the PMM assumes a Brownian motion process; the POUMM assumes an Ornstein–Uhlenbeck process (Uhlenbeck and Ornstein 1930; Lande 1976; Hansen 1997).

Using the maximum likelihood estimates of the model parameters (supplementary table S1, Supplementary Material online), we simulated random trait trajectories on the UK tree, running 100 replications for each model. For each replication, we estimated the correlation, r_{A}, in PPs using the simulated values instead of the real values. The resulting correlation estimates are shown on figure 3 as brown and green points and error bars for the PMM and POUMM simulations, respectively. We notice that there is a significant difference between the correlation estimates of the two models. In particular, in the leftmost decile the POUMM estimate is significantly higher than the PMM estimate (the POUMM 95% CI excludes the PMM estimate).

In order to understand the above difference between PMM and POUMM, we derive approximate analytical expressions of the correlation as a function of d_{ij} under the two models. Assume for simplicity that two tips i and j are situated at equal distance, t, from the root. According to Brownian motion (BM), the correlation is a function of t and the

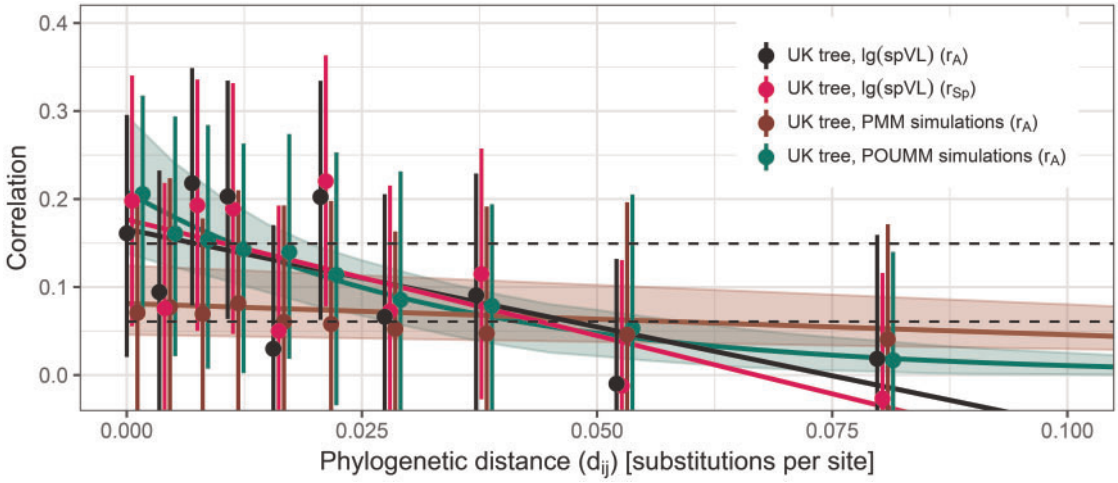


Fig. 3. Correlation between lg(spVL)-values in HIV phylogenetic pairs. A sample of 1917 PPs with lg(spVL)-measurements from HIV patients shows a decrease in the correlation (ICC) between pair trait values as a function of the pair phylogenetic distance d_{ij}. The point estimates and 95% CIs in ten strata of equal size (deciles) are depicted as points and error bars positioned at the mean d_{ij} for each stratum, $\bar{d}_{ij}$. Black and magenta points with error-bars denote the estimated r_A and r_{Sp} in the real data. Dashed horizontal bars denote the 95% CI for r_A evaluated on all phylogenetic pairs. A black and a magenta inclined line denote the least squares linear regression of r_A and r_{Sp} on $\bar{d}_{ij}$. Brown and green points with error bars denote the estimated values of r_A obtained after replacing the real trait values on the tree by values simulated under the maximum likelihood fit of the PMM and the POUMM methods, respectively (mean and 95% CI estimated from 100 replications). A brown and a green line show the expected correlation between pairs of tips at distance d_{ij}, as modeled under the ML-fit of the PMM and the POUMM (eqs. 2 and 3). A light-brown and a light-green region depict the 95% high posterior density (HPD) intervals inferred from Bayesian fit of the two models (Materials and Methods).

distance t_{ij} from the root to the pair's most recent common ancestor (mrca):

$$\begin{aligned} r_{BM,ij} &= \frac{\mathrm{Cov}_{BM}(t_{ij};\sigma^2)}{\mathrm{Var}_{BM}(t;\sigma^2)+\sigma_e^2} \\ &= \frac{\sigma^2 t_{ij}}{\sigma^2 t+\sigma_e^2}, \end{aligned} \tag{2}$$

where σ^2 denotes the unit time variance of the BM process and σ_e^2 denotes the variance of the environmental (nonheritable) component, e (Housworth et al. 2004, Materials and Methods). According to Ornstein–Uhlenbeck (OU), the correlation is a function of t, t_{ij}, as well as the phylogenetic distance between the tips, d_{ij}:

$$\begin{aligned} r_{OU,(ij)} &= \frac{\mathrm{Cov}_{OU}(t_{ij},d_{ij};\alpha,\sigma^2)}{\mathrm{Var}_{OU}(t;\alpha,\sigma^2)+\sigma_e^2} \\ &= \frac{\frac{\sigma^2}{2\alpha}\exp(-\alpha d_{ij})\left(1-\exp(-2\alpha t_{ij})\right)}{\frac{\sigma^2}{2\alpha}\left(1-\exp(-2\alpha t)\right)+\sigma_e^2}, \end{aligned} \tag{3}$$

where the additional parameter α denotes the selection strength of the OU process (Hansen 1997). By plugging-in the ML estimates for the model parameters (supplementary table S1, Supplementary Material online), substituting t with the mean root-tip distance in the tree ($\bar{t}=0.14$), and approximating t_{ij} with its linear regression on d_{ij} in the UK tree ($\hat{t}_{ij}=0.15-0.63d_{ij}$), we obtain:

$$r_{BM,ij} \approx 0.08 - 0.36 d_{ij}. \tag{4}$$

$$\begin{aligned} r_{OU,(ij)} &\approx 0.21\exp(-28.78 d_{ij}) \\ &\times\left(1-\underbrace{\exp(-8.35+36.47 d_{ij})}_{\approx 0}\right) \\ &\approx 0.21\exp(-28.78 d_{ij}). \end{aligned} \tag{5}$$

The last approximation in equation (5) follows from the fact that the term $\exp(-8.35+36.47 d_{ij})$ is nearly 0 for the range of phylogenetic distances ($d_{ij}\in[0,0.14]$) in the UK tree (see supplementary information, Supplementary Material online, for further details on the above approximations).

Equations (4) and (5) represent a linear and an exponential model of the correlation as a function of d_{ij}. The values of these equations at d_{ij}=0 are equal to the phylogenetic heritabilities estimated at the mean root-tip distance $\bar{t}$ under PMM and POUMM (details on that later). The slope of the linear model (eq. 4) equals −0.36 (95% HPD [−0.58, −0.21]). The rate of the exponential decay (eq. 5) equals the POUMM parameter α=28.78 (95% HPD [16.64, 46.93]) and the half-life of decay equals $\ln(2)/\alpha = 0.02$ substitutions per site (95% HPD [0.01, 0.04]).

Plotting the values of equations (4) and (5) and their 95% HPD intervals on figure 3 reveals visually that the POUMM fits better to the data than the PMM. Statistically, this is confirmed by a lower Akaike Information Criterion (AICc) for the POUMM fit and a strictly positive HPD interval for the OU parameter α (supplementary table S1 and fig. S8, Supplementary Material online). The slope of the linear model derived from the PMM fit (eq. 4, brown line on fig. 3) is nearly flat compared with the slopes of the two OLS fits (black and magenta lines on fig. 3). To explain this,

we notice that in PMM, the covariance in phylogenetic pairs and the variance at the population level are modeled as linear functions of the root-mrca distance (t_{ij}) and the root-tip distance (t) (numerator and denominator in eq. 2). Importantly, both of these linear functions are bound to the same slope parameter, σ^2. As it turns out, in the UK data, the covariance and the variance increase at different rates with respect to t_{ij} and t (see supplementary fig. S2 and supplementary information, Supplementary Material online). We conclude that the PMM is not an appropriate model for the correlation in phylogenetic pairs, being unable to model the above difference in the rates.

In the limit $d_{ij} \to 0$, a phylogenetic pair should be equivalent to a DR couple at the moment of transmission, that is, before the genotypes in the two hosts have diverged due to within-host evolution. Thus, it appears reasonable to use an estimate of the correlation at d_{ij}=0 as a proxy for the broad-sense heritability, H^2, in the entire population. This idea has been applied in previous studies of HIV (Hecht et al. 2010; Hollingsworth et al. 2010; Bachmann et al. 2017; Blanquart et al. 2017) as well as malaria (Anderson et al. 2010). One potential obstacle to this approach is the possibility of introducing a sampling bias by filtering of the data. For example, if the study is on a trait, which evolves toward higher values during the course of infection, patients with lower trait values would tend to be more frequent among the CPPs than in the entire population. Thus, there is no guarantee that the trait distribution and, therefore, the heritability measured in the CPPs equals the heritability in the entire population. This problem of sampling bias affects both, resemblance-based as well as the currently used phylogenetic comparative methods. This suggests that the approach of imposing a threshold on d_{ij} or estimating the correlation (r_A, r_{Sp} or another correlation measure) at d_{ij}=0 needs further validation. In the next subsection, we use simulations of the toy model to show that sampling bias, although present, is comparatively small with respect to the negative bias due to measurement delay.

ANOVA-CPP and POUMM Are the Least Biased Heritability Estimators in Toy-Model Simulations

Here, we use simulations of the toy-model to compare a number of heritability estimators against the known true value of H^2 (measured directly by the coefficient of determination R^2_{adj}). We use the symbol T_{10k} to denote the transmission tree of the first 10,000 diagnosed individuals in a simulation. Below we list the different heritability estimators grouping them by the type of their input:

- *Grouping of the trait values by identical pathogen genotype.* We evaluated the coefficient of determination adjusted for finite sample size, R^2_{adj}, and the intraclass correlation (ICC) estimated using one-way ANOVA, $r_A[\mathrm{id}]$. The main difference between these two estimators is the ANOVA assumption that the group-means (genotypic values) are sampled from a distribution of potentially many more genotypes than the ones found in the data. In contrast, R^2_{adj} assumes that all genotypes in the population are present in the sample. Since the latter assumption is true for the simulated epidemics, R^2_{adj} represents the reference (true) value of H^2 to which all other estimates are compared.
- *Known DR couples.* We evaluated the regression slope of recipient on donor values in three ways: 1) b—based on the trait values at the moment of diagnosing the infection; 2) b_0—based on the trait values right after the transmission events; and 3) $b_{d_{ij}'}$—based on the subsample of diagnosed couples having d_{ij} not exceeding a threshold d_{ij}'. Based on a trade-off between precision and bias, we specified $d_{ij}' = D_1$, D_1 denoting the first decile in the empirical distribution of d_{ij} (see supplementary information, Supplementary Material online).
- *Phylogenetic pairs (PPs) in T_{10k}.* We evaluated ICC using ANOVA in three ways: 1) r_A—based on all PPs; 2) r_{A,D_1}—based on CPPs defined as PPs in T_{10k} having d_{ij} not exceeding the first decile, D_1; and 3) $r_{A,0,\mathrm{lin}}$—the estimated intercept from a linear regression of the values r_{A,D_k} on the mean values $d_{ij,k}$ in each decile, $k = 1, \ldots, 10$; For the latter two estimators, which attempt to estimate r_A at d_{ij}=0, we use the acronym ANOVA-CPP. As an alternative to ANOVA, which is more robust to outliers (e.g., extreme values at the tails of the trait distribution), we evaluated the Spearman correlation in the first decile, hereby denoted as r_{Sp,D_1}.
- *Transmission tree T_{10k}.* We evaluated the phylogenetic heritability based on the ML fit of the PMM and POUMM models. Specifically, we compared the classical formula evaluated at the mean root-tip distance $\bar{t}$ in the tree (eqs. 10 and 12) (Housworth et al. 2004; Leventhal and Bonhoeffer 2016) and the empirical formula based on the sample trait variance, $s^2(z)$ (eqs. 11 and 13) (described in Materials and Methods). For the PMM, we denote these estimators by $H^2_{BM}(\bar{t})$ and H^2_{BMe}; for the POUMM, we use the symbols $H^2_{OU}(\bar{t})$ and H^2_{OUe}:

Table 1 summarizes the mathematical definition and the assumptions of the above estimators. A more detailed description of the PMM and the POUMM methods is provided in Materials and Methods. The referenced textbooks on quantitative genetics (Lynch and Walsh 1998) are excellent references for the other methods.

By combining "neutral" and "select" dynamics for the strain mutation and substitution rates at the within-host level, and the virus-induced per capita death rate and per contact transmission probability at the between-host level, we defined the following scenarios of the toy-model:

- Within: neutral/Between: neutral;
- Within: select/Between: neutral;
- Within: neutral/Between: select;
- Within: select/Between: select;

For each of these scenarios and mean contact interval $1/\kappa \in \{2, 4, 6, 8, 10, 12\}$ (arbitrary time units), we executed ten simulations resulting in a total of $4 \times 6 \times 10 = 240$ simulations. Of the 240 simulations, 175 resulted in epidemic outbreaks of at least 10,000 diagnosed hosts. For each

Table 1. Tested Estimators of the Broad-Sense Heritability of Pathogen Traits.

Input Data	Method (Abbreviation)	Assumptions	Estimator
Grouping by identical infecting strain	Adjusted coefficient of determination	The sample of data contains all genotypes present in the population	$R^2_{adj} = 1 - \frac{N-1}{N-K}\frac{s^2(z-\hat{G})}{s^2(z)}$ (6)
	One-way analysis of variance (ANOVA)	Independently sampled genotypes i.i.n.d. trait-values within each group Equal within-group variances (homoscedasticity)	$r_A[id] = \frac{(M_{Sb}-M_{Se})/n}{(M_{Sb}-M_{Se})/n+M_{Se}}$ (7)
Known donor–recipient couples	Donor–recipient regression (DR)	Independently sampled donor–recipient couples Equal residual variance across the range of donor-values (homoscedasticity) Equal donor and population variances	$b = \frac{s(z_{don},z_{rcp})}{s^2(z_{don})}$, (8) variants: b, b_0, $b_{d_{ij}'}$
Phylogenetic pairs (PPs)	ANOVA on all/closest PPs (ANOVA-PP, ANOVA-CPP)	ANOVA assumptions (see above)	Defined as in equation (7), but calculated on PPs variants: r_A, $r_{A,d_{ij}'}$
	Spearman correlation on all/closest PPs	PPs are independent from one another	Pearson (product mean) correlation, calculated on the ranks of the trait-values. variants: r_{Sp}, $r_{Sp,d_{ij}'}$
	Linear regression of r_A on d_{ij} upon a stratification	$r_A$depends linearly on d_{ij} Equal residual variance across the range of d_{ij}	The intercept, $r_{A,0,lin}$, from the OLS fit of the model $r_A(d_{ij}) = r_{A,0,lin} + \omega_1 d_{ij}$. (9)
Transmission tree	Phylogenetic mixed model (PMM)	Branching BM evolution	$H^2_{BM}(\bar{t}) = \bar{t}\sigma^2/(\bar{t}\sigma^2+\sigma_e^2)$ (10)
		i.i.n.d. distributed environmental deviation, $e \sim N(0,\sigma_e^2)$	$H^2_{BMe} = 1 - \sigma_e^2/s^2(z)$(11)
	Phylogenetic Ornstein–Uhlenbeck mixed model (POUMM)	Branching OU evolution	$H^2_{OU}(\bar{t}) = \frac{\sigma^2(1-\exp(-2\alpha\bar{t}))}{\sigma^2(1-\exp(-2\alpha\bar{t}))+2\alpha\sigma_e^2}$ (12)
		i.i.n.d. environmental deviation, $e \sim N(0,\sigma_e^2)$	$H^2_{OUe} = 1 - \sigma_e^2/s^2(z)$(13)

NOTE.—Notation: $s^2(\cdot)$, sample variance; $s(\cdot,\cdot)$, sample covariance; N, number of patients; K, number of distinct groups of patients, that is, genotypes or phylogenetic pairs; z, measured values; $\hat{G}$, estimated genotypic values: mean values from patients carrying a given genotype; z_{don}, donor values; z_{rcp}, recipient values; M_{Se}, within-group mean square: $M_{Se} = \frac{\sum(z_i-\bar{z}_i)^2}{N-K}$, where z_i is an individual's value and $\bar{z}_i$ is the mean value of the group to which the individual belongs; M_{Sb}, among-group mean square: $M_{Sb} = \frac{\sum(\bar{z}_i-\bar{z})^2}{K-1}$, where $\bar{z}_i$ is defined as above and $\bar{z}$ is the population mean value; n, weighted mean number patients in a group, that is, n=2 for phylogenetic pairs and $n = \left(N - \frac{\sum n_i^2}{N}\right)/(K-1)$ for groups of variable size; α, σ, σ_e: PMM/POUMM parameters (described in Materials and Methods).
i.i.n.d., independent and identically normally distributed; d_{ij}, phylogenetic distance between donor–recipient pairs or phylogenetic pairs; d_{ij}', threshold on d_{ij} (see text).

outbreak, we analyzed the populations of the first up to 10,000 diagnosed hosts.

Rarer transmission events (bigger $1/\kappa$) result in longer transmission trees and, therefore, longer average phylogenetic distance between tips, d_{ij} (supplementary fig. S3, Supplementary Material online). This enabled demonstrating the effect of accumulating within-host evolution on the different heritability estimators (fig. 4).

Figure 4 shows that the estimators b_{D_1}, b, r_{A,D_1}, and r_A are negatively biased in general for all toy-model scenarios. This bias tends to increase with the mean contact interval, $1/\kappa$ (respectively, d_{ij}), because random within-host mutation tends to decrease the genetic overlap between DRs and phylogenetic pairs (supplementary fig. S4, Supplementary Material online). The negative bias was far less pronounced when imposing a threshold on d_{ij} but this came at the cost of precision (less biased but longer box-whisker plots for b_{D_1} and r_{A,D_1} compared with b and r_A) (fig. 4). Several additional sources of bias were revealed when considering the practically unavailable estimators b_0 and $r_A[id]$. The estimator $r_A[id]$ was positively biased due to the small number of simulated genotypes (only six)—this was validated through additional simulations showing that $r_A[id]$ converges to the true value for a slightly bigger number of genotypes (e.g., $K \geq 24$ genotypes, see supplementary information, Supplementary Material online). The estimator b_0 was behaving accurately in the neutral/neutral scenario (excluding very short contact intervals) but tended to have a bias in both directions in all scenarios involving selection. The main reason for these biases was the phenomenon of "sampling bias" consisting in a difference between the distributions of measured values in the DR couples and the population of interest. Although its magnitude was comparatevely small in the simulations, we presume that sampling bias could play an important role in real biological applications. We already gave an example of this bias in the previous subsection. Another manifestation of sampling bias

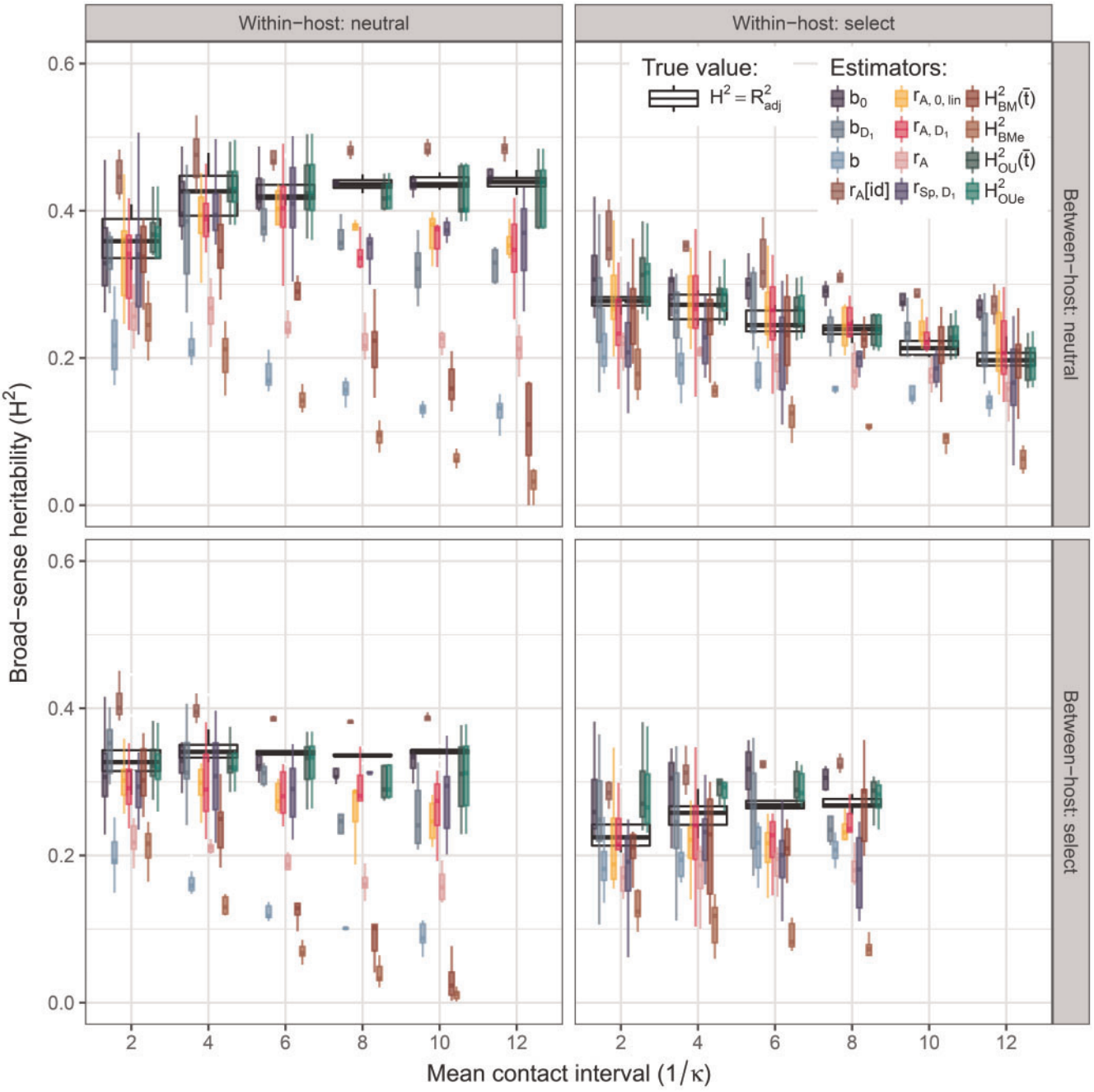


FIG. 4. Heritability estimates in toy-model simulations. (*A*–*D*) H^2-estimates in simulations of "neutral" and "select" within-/between-host dynamics. Each group of box-whiskers summarizes the simulations for a fixed scenario and contact interval, $1/\kappa$; white boxes (background) denote true heritability, colored boxes denote estimates (foreground). Statistical significance is evaluated through *t*-tests summarized in table 2.

is the fact that b_0 does not fully eliminate the effect of within host-evolution (and selection) in the donors. This is why, in cases of selection, the phenotypic variance in the donors tends to be smaller than the variance in the recipients as well as the variance in the population (supplementary fig. S5, Supplementary Material online). Additional details on these potential sources of bias are provided in supplementary information, Supplementary Material online.

Further, the simulations showed that a worsening fit of the BM model on longer transmission trees was causing an inflated estimate of the environmental deviation, σ_e, in the PMM fits and, therefore, a negative bias in $H^2_{BM}(\bar{t})$ and H^2_{BMe} (compare estimates for small and big values of $1/\kappa$ on fig. 4 and supplementary fig. S6*C*, Supplementary Material online). In contrast with the PMM, the POUMM estimates, $H^2_{OU}(\bar{t})$ and H^2_{OUe} were far more accurate and the value of σ_e in the POUMM ML fit was nearly matching the true nonheritable deviation in most simulations (fig. 4 and supplementary fig. S6*C*, Supplementary Material online). The better ML fit of the POUMM was confirmed by stronger statistical support, namely by lower AICc values in all toy-model simulations (supplementary fig. S6*D*, Supplementary Material online).

The fact that the POUMM outperformed the PMM in all scenarios contradicted with the initial belief that the PMM should be the better suited model for a neutrally evolving trait represented by the neutral/neutral scenario, whereas the POUMM should fit better to scenarios involving selection. It was also counterintuitive that the inferred parameter α from the POUMM model was significantly positive in all simulations including the neutral/neutral scenario (supplementary fig. S6*B*, Supplementary Material online). To better understand this phenomenon, we performed the PP stratification analysis on the toy-model data (supplementary fig. S7, Supplementary Material online). This revealed a pattern of correlation that decays exponentially with d_{ij}. The shape of exponential decay was mostly pronounced for longer contact intervals, $1/\kappa$, particularly in the neutral/neutral scenario (first column on supplementary fig. S7, Supplementary Material online). In supplementary information, Supplementary Material online, we show that an exponentially decaying phenotypic correlation is consistent with a neutrally mutating genotype under a Jukes–Cantor substitution model (Yang 2006). The decay of the correlation was still present in scenarios involving within- and/or between-host selection but the observed pattern was rather irregular and deviating from an exponential function of d_{ij} (supplementary fig. S7, Supplementary Material online). In most cases, the ML fit of the PMM method was a bad fit to the decay of correlation (brown dots and error-bars on supplementary fig. S7, Supplementary Material online); for longer contact intervals, there was a tendency toward constant values of the correlation under PMM far below the true value (brown dots and error bars on supplementary fig. S7, Supplementary Material online). This explains the overall better accuracy of the POUMM versus the PMM method.

Table 2 shows the average bias of each tested estimator for each of the four scenarios. We conclude that, apart from the practically inaccessible estimators based on grouping by identical genotype (R^2_{adj} and $r_A[\text{id}]$), the most accurate estimators of H^2 in the toy-model simulations are $H^2_{\text{OU}}(\bar{t})$ and $H^2_{\text{OU}e}$ followed by estimators of the correlation in PPs minimizing the phylogenetic distance d_{ij}, that is (r_{A,D_1}, $r_{A,0,lin}$, r_{Sp,D_1}). In the next subsection, we report the results from these estimators in the UK HIV data.

Table 2. Mean Difference $\widehat{H^2} - R^2_{\text{adj}}$ from the Toy-Model Simulations Grouped by Scenario.

Within: / Between:	Neutral / Neutral	Neutral / Select	Select / Neutral	Select / Select
N	50	41	47	37
b_0	−0.01*	−0.02**	0.05**	0.04**
b_{D_1}	−0.07**	−0.04**	0	−0.01
b	−0.25**	−0.2**	−0.07**	−0.06**
$r_A[id]$	0.05**	0.05**	0.08**	0.06**
$r_{A,\hat{0},lin}$	−0.05**	−0.06**	0.01	−0.04**
r_{A,D_1}	−0.05**	−0.06**	0	−0.03*
r_A	−0.18**	−0.15**	−0.06**	−0.08**
r_{Sp,D_1}	−0.05**	−0.05**	−0.05**	−0.07**
$H^2_{\text{BM}}(\bar{t})$	−0.17**	−0.17**	−0.01	−0.04*
$H^2_{\text{BM}e}$	−0.28**	−0.24**	−0.12**	−0.16**
$H^2_{\text{OU}}(\bar{t})$	−0.01	−0.02**	0.01*	0.03**
$H^2_{\text{OU}e}$	−0.01	−0.02**	0.01*	0.03**

NOTE.—Statistical significance is estimated by Student's *t*-tests, *P* values denoted by an asterisk as follows: * $P<0.01$; ** $P<0.001$. Gray background indicates estimates that are unavailable in practice.

Heratibality of lg(spVL) in the UK HIV Cohort

We evaluated the correlation in the CPPs (ANOVA and Spearman correlation) in data from the UK HIV cohort comprising lg(spVL) measurements and a tree of viral (*pol*) sequences from 8,483 patients inferred previously in (Hodcroft et al. 2014). In addition, we performed a Bayesian fit of the POUMM and the PMM methods to the same data. The goal was to test our conclusions on a real data set and to compare the H^2-estimates from CPPs and POUMM to previous PMM/ReML-estimates on exactly the same data (Hodcroft et al. 2014).

In applying ANOVA-CPP, the first step has been to define the threshold phylogenetic distance for defining CPPs. To that end, we explored different stratifications of the PPs as shown on supplementary figure S1*B*, Supplementary Material online, and a scatter plot of the phylogenetic distances against the absolute phenotypic differences, $|\Delta \lg(\text{spVL})|$ (fig. 5*A*). This revealed a small set of 116 PPs having $d_{ij} \leq 10^{-4}$ and narrowly coinciding with the first vigintile (also called 20-quantile or ventile) of d_{ij}. The phylogenetic distance in all remaining tip-pairs was more than an order of magnitude bigger, that is, $d_{ij} > 10^{-3}$. Given that the phylogenetic distance on the transmission tree is measured in substitutions per site and the length of the *pol*-region is in the order of 10^3 sites, we presume that the above set of 116 PPs corresponds to a set of 116 pairs of identical *pol* consensus sequences (no sequence data were available to check this). Based on this observation, we defined the above pairs as CPPs and the threshold was formally set to $d_{ij}' = 10^{-4}$. We validated that the CPPs were randomly distributed along the tree (fig. 5*B*). The random distribution of the CPPs along the transmission tree suggests that these phylogenetic pairs correspond to randomly occurring early detections of infection (trait values from each pair depicted as magenta segments on fig. 5*B*). To check that the filtering of the data, did not introduce a considerable sampling bias due to selection (see previous subsection), we also validated that there was no substantial difference in the trait distributions of all patients, the PPs and the CPPs (fig. 5*C*).

We compared the following estimators of H^2:

- ANOVA-CPPs (r_{A,D_1}, $r_{A,10^{-4}}$, r_{A,V_1}) and the original PP-method r_A;
- The intercept from the linear regression of r_A on d_{ij} upon a stratification of the PPs into deciles ($r_{A,0,lin}$, eq. 9);
- Spearman correlatoin in CPPs (r_{Sp,D_1}, $r_{\text{Sp},10^{-4}}$, r_{Sp,V_1}) and in all PPs (r_{Sp}).
- The intercept from the linear regression of r_{Sp} on d_{ij} upon a stratification of the PPs into deciles ($r_{\text{Sp},0,lin}$);
- POUMM ($H^2_{\text{OU}}(\bar{t})$, $H^2_{\text{OU}e}$), versus PMM ($H^2_{\text{BM}}(\bar{t})$, $H^2_{\text{BM}e}$) on the entire tree;

The results from these analyses are reported in table 3. ANOVA- and Spearman-correlation estimates, which

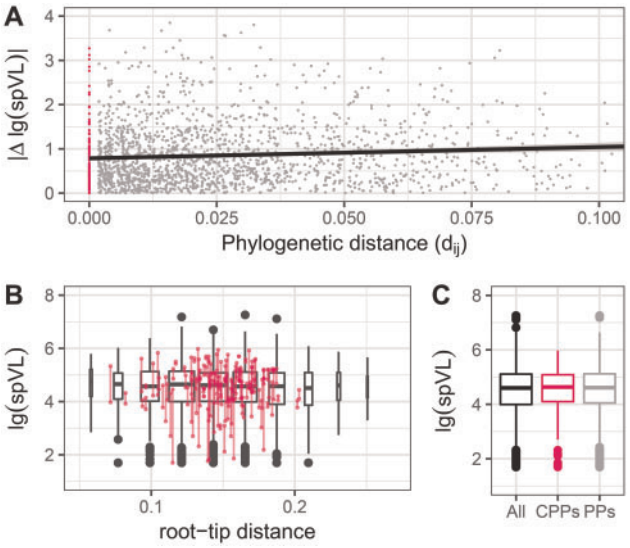


FIG. 5. Phylogenetic pairs in lg(spVL) data from the United Kingdom. (*A*) A scatter plot of the phylogenetic distances between pairs of tips against their absolute phenotypic differences: gray, PPs ($d_{ij} > 10^{-4}$); magenta, CPPs ($d_{ij} < 10^{-4}$). A black line shows the linear regression of $|\Delta \lg(\text{spVL})|$ on d_{ij} (the slope of the regression was statistically positive at the 0.01 level). (*B*) A box-plot representing the trait-distribution along the transmission tree. Each box-whisker represents the lg(spVL)-distribution of patients grouped by their distance from the root of the tree measured in substitutions per site. Wider boxes indicate groups bigger in size. Segments in magenta denote lg(spVL)-values in CPPs. (*C*) A box-plot of the lg(spVL)-distribution in all patients (black), PPs (gray), and CPPs (magenta).

minimized the phylogenetic distance by means of regression or filtering of the phylogenetic pairs had point-estimates of $r_{A,10^{-4}} = 0.17$ and $r_{Sp,10^{-4}} = 0.22$. The slightly higher estimate for the Spearman correlation could be explained by the presence of outliers in the data. Applying the POUMM to the entire tree reported a point estimate $H^2_{OU}(\bar{t}) = 0.21$ (8,483 patients, 95% CI [0.14, 0.29]).

Conversely, the heritability estimates from the original PP method (ANOVA or Spearman correlatoin on all PPs) and the PMM were significantly lower and falling below the 95% CIs from the POUMM (table 3). This confirms the observation from the toy-model simulations that these estimators are negatively biased, since they ignore or inaccurately model the changing correlation within pairs of tips. We validated the stronger statistical support for the POUMM with respect to the PMM, by its lower AICc value (supplementary table S1, Supplementary Material online) and by the posterior density for the POUMM parameter α (supplementary fig. S8, Supplementary Material online).

Finally, we compared our estimates of lg(spVL)-heritability to previous applications of the same methods on different data sets (fig. 6). In agreement with the toy-model simulations, estimates of H^2 using PMM or other BM-based phylogenetic methods (i.e., Blomberg's K and Pagel's λ) are notably lower than all other estimates, suggesting that these phylogenetic comparative methods underestimate H^2; resemblance-based estimates are down-biased by measurement delays (e.g., compare early vs. late in the Netherlands on fig. 6).

In summary, POUMM and ANOVA-CPP yield agreeing estimates for H^2 in the UK data and these estimates agree with resemblance-based estimates in data sets with short measurement delay (different African countries and the Netherlands). Similar to the toy-model simulations, we notice a well-pronounced pattern of negative bias for the other estimators, PMM and ANOVA-PP, as well as for the previous resemblance-based studies on data with long measurement delay.

Discussion

Clarifying the Terminology and Notation

In this study, we explored how the differences between pathogens and mating species affect the various tools employed in estimating the heritability of pathogen traits. For mating species, the resemblance between relatives has been directly associated with the genetic determination of quantitative traits. The most prominent example is the parent–offspring regression slope used to estimate the narrow-sense heritability, h^2. For pathogens, one needs to disentangle the concepts of resemblance and genetic determination. First of all, the only reason to associate the parent–offspring regression slope with narrow-sense heritability is the presence of genetic segregation and recombination during sexual reproduction, favoring the inheritance of single-locus additive effects over multilocus epistatic effects (Lynch and Walsh 1998). Given that clonal pathogen transmission excludes segregation and recombination, the above association is invalid for pathogen traits. The correlation between transmission partners should rather be associated with the broad-sense heritability, H^2. This association, though, is compromized by a number of sources of bias, such as partial quasispecies transmission, within-host evolution, and many potential cofactors, such as shared habitats between donors and recipients, sampling bias, and convergent within-host evolution. All methods reviewed in this article can be regarded as methods that estimate the correlation between patients infected with identical pathogen strains. This is true also for the phylogenetic approaches, since, technically, the phylogenetic heritability is the expected correlation between pairs of tips in the limit $d_{ij} \to 0$ (see also Materials and Methods). Thus, all estimators can only be regarded as statistics summarizing the resemblance that is still observable in the presence of the above factors.

A Disagreement between Simulation Studies

Using simulations of the toy epidemiological model, we have shown that two methods based on phenotypic and sequence data from patients—estimating the correlation in CPPs and fitting the POUMM to the data—provide more accurate heritability estimates compared with previous approaches like DR and PMM. However, we should not neglect the arising discrepancy between our and previous simulation reports advocating either PMM (Hodcroft et al. 2014) or DR (Leventhal and Bonhoeffer 2016) as unbiased heritability

Table 3. Estimates of lg(spVL)-Heritability in HIV Data from the United Kingdom.

Method	*N*	$\hat{H}^2$	95% CI	95% HPD
Linear regression of r_A on $\bar{d}_{ij}$ in deciles (eq. 9) ($r_{A,0,lin}$)	10 points	0.17	[0.09, 0.24]	–
Linear regression of r_{Sp} on $\bar{d}_{ij}$ in deciles ($r_{Sp,0,lin}$)	10 points	0.18	[0.11, 0.25]	–
ANOVA-CPP (r_{A,V_1})	224	0.17	[−0.02, 0.31]	–
ANOVA-CPP ($r_{A,10^{-4}}$)	232	0.16	[0.01, 0.30]	–
ANOVA-CPP (r_{A,D_1})	384	0.16	[0.06, 0.25]	–
ANOVA-PP (r_A)[a]	3,834	0.11	[0.08, 0.14]	–
Spearman-CPP (r_{Sp,V_1})	224	0.23	[0.05, 0.42]	–
Spearman-CPP ($r_{Sp,10^{-4}}$)	232	0.22	[0.03, 0.4]	–
Spearman-CPP (r_{Sp,D_1})	384	0.2	[0.06, 0.34]	–
Spearman-PP (r_{Sp})[a]	3,834	0.11	[0.06, 0.15]	–
POUMM ($H^2_{OU}(\bar{t})$)	8,483	0.21	–	[0.14, 0.29]
POUMM (H^2_{OUe})	8,483	0.2	–	[0.13, 0.29]
PMM ($H^2_{BM}(\bar{t})$)[b]	8,483	0.08	–	[0.05, 0.12]
PMM (H^2_{BMe})[b]	8,483	0.06	–	[0.02, 0.1]
PMM, ReML (Hodcroft et al. 2014)[b]	8,483	0.06	[0.03, 0.09]	–

NOTE.—Also written are the results from a previous analysis on the same data set (Hodcroft et al. 2014). "–": the analysis was not done in the mentioned study. Gray background: estimates considered unreliable due to: [a]negative bias caused by measurement delays and [b]negative bias caused by BM violation. Uncertainty in the estimates is expressed in terms of 95% confidence intervals (CI), or, in the case of Bayesian inference, by 95% high posterior density intervals (HPDs).

estimators. Both of these studies have modeled within-host evolution, but failed to demonstrate the biases shown in this article. This could be explained by simulation artifacts. Hodcroft et al. (2014) perform simulations under a PMM model, so it is unlikely to reveal any bias in the PMM estimator; Leventhal and Bonhoeffer (2016) evaluated DR in consecutive Wright–Fisher generations using the donor values at the moment of transmission, thus, excluding potential measurement delay in the donors and accounting for a minute measurement delay in the recipients (one generation on the scale of hundreds of simulated generations). Compared with these simulations, the toy-model presented here has several important advantages: 1) it is biologically motivated by phenomena such as pathogen sequence mutation during infection, transmission of entire pathogens instead of proportions of trait values, and within-/between-host selection; 2) it allows to compare various resemblance-based and phylogenetic heritability estimates against the direct estimator, R^2_{adj}; 3) it is a fair test for all estimators of heritability, because it does not obey any of the estimators' assumptions, such as linearity of recipient—on donor values, normality of trait values, OU or BM evolution, independence between pathogen and host effects; and 4) it generates transmission trees that reflect the between-host dynamics, for example, clades with higher trait values exhibit denser branching in cases of between-host selection. As a criticism, we note that the toy-model does not allow strain coexistence within a host and, thus, is not able to model partial quasispecies transmission and, in particular, transmission bottlenecks (Keele et al. 2008) or preferential transmission of founder strains (Lythgoe and Fraser 2012). Although it may be exciting from a biological point of view, the inclusion of strain coexistence comes with a series of conceptual challenges, such as the definition of genotype and clonal identity or the formulation of the trait value as a function of a quasispecies—instead of a single strain genotype. These challenges should be addressed in future studies implementing more advanced models of within-host dynamics and leveraging deep sequencing data. To conclude, the discrepancy between simulation studies highlights that no inference method suits all simulation setups *ergo* biological contexts. Thus, rather than proving universality of a particular method, simulations should be used primarily to study how particular biologically relevant features affect the methods on the table.

The Heritability of HIV Set-Point Viral Load Is at Least 20%

Applied to data from the United Kingdom, POUMM reported three times higher point estimates and nonoverlapping HPDs compared with a previous PMM/ReML-based estimate on the same data (0.06, 95% CI [0.02, 0.09]) (Hodcroft et al. 2014). Our PMM implementation confirmed this estimate. However, based on figure 3 and our simulations (fig. 4), the PMM estimates are underestimates of the true heritability. The estimate of 20% should still be considered a lower bound since it does not account for additional sources of potential negative bias, such as partial quasispecies transmission and measurement error. This result matches estimates from GWAS studies on the pathogen revealing that genetic polymorphisms in the virus explain ∼20% from spVL variance in other cohorts (reviewed in Bonhoeffer et al. 2015). Overall, our analyses yield an unprecedented agreement between estimates of DR resemblance and phylogenetic heritability in large European data sets and African cohorts, provided that measurements with large delays have been filtered out prior to resemblance evaluation (Hecht et al. 2010; Hollingsworth et al. 2010) (fig. 6*A*). Also noteworthy are the facts that our estimates for the UK data set support the results from Fraser et al. (2014) who conducted a meta-analysis of three data sets on known transmission partners (Hollingsworth et al. 2010; Lingappa et al. 2013; Yue et al. 2013) (433 pairs in total) reporting heritability values of

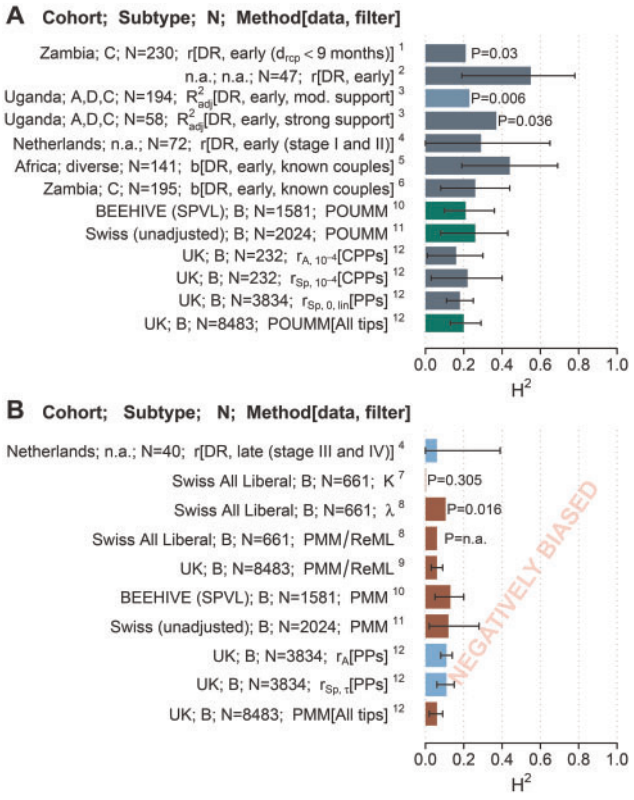


Fig. 6. A comparison between H^2H^2-estimates from the UK HIV-cohort and previous estimates on African, Swiss, and Dutch data. (*A*) Estimates with minimized measurement delay (dark cadet-blue) and POUMM estimates (green); (*B*) Down-biased estimates due to higher measurement delays (light-blue) or violated BM-assumption (brown). Confidence is depicted either as segments indicating estimated 95% CI or *P* values in cases of missing 95% CIs. References to the corresponding publications are written as numbers in superscript as follows: 1: Tang et al. (2004); 2: Hecht et al. (2010); 3: Hollingsworth et al. (2010); 4: van der Kuyl et al. (2010); 5: Lingappa et al. (2013); 6: Yue et al. (2013); 7: Alizon et al. (2010); 8: Shirreff et al. (2013); 9: Hodcroft et al. (2014); 10: Blanquart et al. (2017); 11: Bertels et al. (2018); 12: this work. For clarity, estimates from previous studies, which are not directly comparable (e.g., previous results from Swiss MSM/strict data sets; Alizon et al. 2010).

0.33, CI [0.20, 0.46], as well as the recent results from Blanquart et al. (2017) who conducted a POUMM and a PMM analysis on a whole-genome meta-data set (1,581 sequences from several European countries) reporting spVL heritability of 0.31, CI [0.15, 0.43]. In analogy with our ANOVA approach, Blanquart et al. (2017) measured the Pearson correlation in "cherries" partitioned by phylogenetic distance, showing a similar pattern of decreasing correlation with d_{ij}. Contrary to the UK data though, Blanquart et al. (2017) have shown nearly equal statistical support for PMM (α=0, AIC = 3,343.2) and POUMM (α=7.6, 95% bootstrap CI [1.2, 10.0], AIC = 3,344.5) for 1,581 subtype B *pol* sequences and spVL measurements (table 1 in Blanquart et al. 2017). This equal support fot the PMM and the POUMM models might indicate that none of the two models is a good fit to the data (i.e., flat likelihood surface), or that the likelihood surface for the POUMM is bimodal with modes at α=0 and at α=7.6. A Bayesian POUMM fit with uninformative prior could be used to reveal such anomailies (see Materials and Methods and supplementary fig. S8, Supplementary Material online).

To sum up, all data sets support the hypothesis of HIV influencing spVL (H^2>0.2). The particular estimates provided here should be interpreted as lower bounds for H^2, because the partial quasispecies transmission, the noises in spVL measurements and the noise in transmission trees are included implicitly as environmental (nontransmittable) effects. The nonzero heritability motivates further HIV whole-genome sequencing (Metzner 2016) and genome-wide studies of the viral genetic association with viral load and virulence.

A Critical View on the POUMM

The OU process has found previous applications as a model for stabilizing selection in macroevolutionary studies (Lande 1976; Felsenstein 1988; Hansen 1997; Hansen and Bartoszek 2012) and references therein. As a contribution of this work, we have shown that the OU process is well adapted for the modeling of pathogen evolution along transmission trees in both, neutral as well as selection scenarios. The key advantage of the OU process to the BM process is the way in which the phylogenetic distance between a pair of tips enters in the expression for their correlation (eq. 3). This is a crucial advantage in modeling the loss of resemblance caused by within-host evolution of the pathogen (fig. 3 and supplementary fig. S7, Supplementary Material online). But there is a caveat coming along with this property of the OU-model—both, the rate at which a trait evolving under OU adapts toward θ and the rate of correlation decay for a pair of tips are governed by the same parameter: α. This is why a significantly positive estimate for α does not necessarily imply stabilizing selection. This was clearly shown in the neutral/neutral scenario of the toy-model simulations (supplementary fig. S6*B*, Supplementary Material online). A further extension of the POUMM using two separate parameters for the rate of attraction toward θ and for the rate of decorrelation would allow to disentangle the two forces.

Most of the above-mentioned studies and the accompanying software packages implementing phylogenetic OU models have assumed that the whole trait evolves according to an OU process, usually disregarding the presence of a biologically relevant nonheritable component *e* or treating it as a measurement error whose variance is a priori known (FitzJohn 2012). Having the OU process act on the genotypic values rather than whole trait values is a simplifying assumption facilitating mathematical processing (Mitov and Stadler 2016). However, our toy model simulations have shown robustness and statistical power of the POUMM in complicated scenarios combining trait-based selection at the within- and between-host levels.

A last criticism that can be addressed to the POUMM method is that it is unaware of between-host selection and demographic processes, which may result in a correlation between tree structure and trait values (e.g., higher branching density in clades with higher z). As noted by Leventhal and Bonhoeffer (2016), this is a general issue with phylogenetic comparative approaches assuming a global evolutionary

process acting on the whole phylogeny. An unexplored alternative would be to associate different instances of POUMM to different clades in the tree based on prior knowledge about heterogeneity between these clades.

Outlook

ANOVA-CPP and POUMM have great potential to become widely used tools in the study of pathogens. The accompanying R-package patherit provides a common interface for using the two methods on a transmission tree and phenotype data (Materials and Methods). ANOVA-CPP works on pairs of trait values from carriers of nearly identical strains and can be easily extended to groups of variable size (Lynch and Walsh 1998; Anderson et al. 2010). Thus, ANOVA-CPP is ideal for slowly evolving pathogens such as DNA-viruses, bacteria, and protozoa, where clusters of patients carrying identical-by-descent (IBD) strains are frequently found. For example, Anderson et al. (2010) identified 27 clusters of two to eight carriers of IBD strains in a small set of 185 malaria patients, that is, 41% of the patients participated in clusters. On the other hand, IBD-pairs are rare for rapidly evolving RNA-viruses, such as HIV and HCV. For instance, we identified only 116 CPPs in a large data set of 8,483 HIV-sequences, that is, <3% of the patients involved in IBD-pairs. However, the rapidly accumulating sequence diversity of RNA-viruses allows building large-scale phylogenies, which approximate transmission trees between patients. Thus, RNA-viruses should make the ideal scope for the POUMM. If the transmission tree is large enough, it is be possible to compare the estimates from the two methods and to analyze the profile of the correlation in phylogenetic pairs, as we did in the UK HIV data (fig. 3 and supplementary fig. S2, Supplementary Material online). We believe that, together, the two methods enable accurate and robust heritability estimation in a broad range of pathogens.

Materials and Methods

The subsections below provide details on the different heritability estimators (based on the categorization by input type, table 1) and the toy-model simulations.

Grouping by Identical Infecting Strain

Adjusted Coefficient of Determination

We calculated R^2_{adj} based on equation (6) (table 1).

One-Way Analysis of Variance

We calculated r_A based on equation (7) (table 1). A more detailed description of one-way ANOVA can be found in chapter 18 of Lynch and Walsh (1998).

Donor–Recipient Couples

To calculate the DR regression slope (b, b_0, b_{D_1}), we used equation (8) (table 1).

Phylogenetic Pairs

To calculate ICC in phylogenetic pairs (r_A, r_{A,D_1}, r_{A,V_1}, $r_{A,10^{-4}}$), we used one-way ANOVA (eq. 7, chapter 18 of Lynch and Walsh 1998). To calculate confidence intervals for the HIV data, we used the R-package "boot" to perform 1,000-replicate bootstraps, upon which we called the package function boot.ci() with type="basic." These confidence intervals were fully contained in the standard ANOVA confidence intervals, based on the F-distribution (Lynch and Walsh 1998), which were slightly wider (not reported).

Phylogenetic Methods

Phylogenetic Mixed Model

The PMM assumes an additive model $z(t) = g(t) + e$, in which $z(t)$ represents the trait value at time t for a given lineage of the tree, $g(t)$ represents a heritable (genotypic) value at time t for this lineage and e represents the environmental (nonheritable) contribution. The genotypic value, $g(t)$, is assumed to evolve according to a branching Brownian motion process defined by the stochastic differential equation:

$$\begin{aligned} dg(t) &= \sigma dW_t, \\ g(0) &= g_0 \end{aligned} \qquad (14)$$

where g_0 is the initial genotypic value at the root, W_t is the standard Wiener process, and $\sigma>0$ is the unit-time SD (Grimmett and Stirzaker 2001).

The environmental contribution e can change along the tree in any way as long as the values e at the tips are independent and identically normally distributed (i.i.n.d.) with mean 0 and variance σ_e^2. In the case of modeling an epidemic, e represents the total contribution from the host immune system, other host factors (e.g., age, sex), the host environment and measurement error; it obtains a value at the beginning of an infection, which can stay constant or change during the course of an infection, but is uncorrelated to the immune system and cofactors of other hosts.

Phylogenetic Ornstein–Uhlenbeck Mixed Model

The POUMM is an extension of the PMM replacing the BM assumption with an assumption of an Ornstein–Uhlenbeck (OU) process for the genotype evolution. The OU-process represents a continuous time random walk, which tends to move around a long-term mean value with greater attraction when the process is further away from that value (Uhlenbeck and Ornstein 1930; Hansen 1997). Technically, this is accomplished by adding an attraction term to equation (14):

$$dg(t) = \underbrace{\alpha[\theta - g(t)]dt}_{\text{Attraction to } \theta} + \underbrace{\sigma dW_t}_{\text{Brownian motion}}, \qquad (15)$$

where θ denotes the long-term mean and $\alpha>0$ is the attraction strength. Since in the limit $\alpha \to 0$ the attraction term vanishes and only the BM term remains, the OU-process represents a generalization of BM. As in the PMM, an independent white noise term $e \sim \mathcal{N}(0, \sigma_e^2)$ is added to $g(t)$ at the tips.

Phylogenetic Heritability

Introduced as a term with the PMM method (Housworth et al. 2004), the phylogenetic heritability quantifies how much

of the trait variance is attributable to g based on a fit of the assumed evolutionary model (in this case, BM or OU). For the BM and the OU processes, the genotypic variance is a function of the model parameters and the time-distance from the root of the ultrametric tree, t (Hansen 1997; Housworth et al. 2004):

$$\mathrm{Var}_{\mathrm{BM}}(t;\,\sigma) = \sigma^2 t \tag{16}$$

$$\mathrm{Var}_{\mathrm{OU}}(t;\,\alpha,\sigma) = \frac{\sigma^2}{2\alpha}(1 - \exp(-2\alpha t)). \tag{17}$$

Given the assumption that g and e are uncorrelated, the phenotypic variance is the sum of the genotypic variance and σ_e^2. Therefore, the phylogenetic heritability is also a function of t:

$$H^2_{\mathrm{BM}}(t;\,\sigma,\sigma_e) = \frac{\mathrm{Var}_{\mathrm{BM}}(t;\,\sigma)}{\mathrm{Var}_{\mathrm{BM}}(t;\,\sigma) + \sigma_e^2} = \frac{\sigma^2 t}{\sigma^2 t + \sigma_e^2}, \tag{18}$$

$$\begin{aligned} H^2_{\mathrm{OU}}(t;\,\alpha,\sigma,\sigma_e) &= \frac{\mathrm{Var}_{\mathrm{OU}}(t;\,\alpha,\sigma)}{\mathrm{Var}_{\mathrm{OU}}(t;\,\alpha,\sigma) + \sigma_e^2} \\ &= \frac{\frac{\sigma^2}{2\alpha}(1 - \exp(-2\alpha t))}{\frac{\sigma^2}{2\alpha}(1 - \exp(-2\alpha t) + \sigma_e^2}. \end{aligned} \tag{19}$$

The above dependency of H^2_{OU} and H^2_{BM} on time is posing a problem in the case of a nonultrametric transmission tree, because the tips are at different time-distance from the root and do not share the same genotypic and phenotypic variance. We tested two possible work arounds: 1) evaluating the heritability at the mean root-tip distance, $\bar{t}$ (Leventhal and Bonhoeffer 2016); and 2) using an empirical definition of the phylogenetic heritability based on the empirical variance in the observed population:

$$H^2_e = 1 - \frac{\sigma_e^2}{s^2(z)}. \tag{20}$$

PMM and POUMM Log-Likelihood

The PMM and the POUMM log-likelihood represents the log-probability density of the observed data at the tips of the tree for given values of the model parameters, $\boldsymbol{\Theta}$. For PMM, $\boldsymbol{\Theta} = <\, g_0, \sigma, \sigma_e\, >$; for POUMM $\boldsymbol{\Theta} = <\, g_0, \alpha, \theta, \sigma, \sigma_e\, >$. Given that the two models are Gaussian, the log-likelihood is defined as the Gaussian log-probability density function:

$$\begin{aligned} \ell\ell(\boldsymbol{\Theta}) = \ln f(\mathbf{z}|\boldsymbol{\Theta}) = -\frac{1}{2}(N\ln(2\pi) + \ln|\mathbf{V}_{\boldsymbol{\Theta}}| + \\ (\mathbf{z} - \boldsymbol{\mu}_{\boldsymbol{\Theta}})'\mathbf{V}_{\boldsymbol{\Theta}}^{-1}(\mathbf{z} - \boldsymbol{\mu}_{\boldsymbol{\Theta}})), \end{aligned} \tag{21}$$

where $\mathbf{z}$ is the observed vector of trait values at the tips, $\boldsymbol{\mu}_{\boldsymbol{\Theta}}$ is the mean vector at the tips ($\mu_i = g_0$ in the case of BM; $\mu_i = \exp(-\alpha t_i)g_0 + (1 - \exp(-\alpha t_i))\theta$ in the case of OU), and $\mathbf{V}_{\boldsymbol{\Theta}}$ is the variance covariance matrix with off-diagonal elements given by the nominators and diagonal elements given by the denominators in equations (2) and (3), respectively.

PMM and POUMM Inference in the Toy-Model Simulations

The POUMM and PMM inference was done using maximum likelihood (ML) fit.

PMM and POUMM Inference on HIV Data

For HIV data, in addition to an ML-fit, we performed a Bayesian (MCMC) fit using an adaptive Metropolis algorithm with coerced acceptance rate (Vihola 2012) written in R (Scheidegger 2012).

The MCMC sampling was performed on the parameters g_0, α, θ, $H^2(\bar{t})$ and σ_e^2 (for likelihood and posterior density calculation, the parameter σ^2 was mapped back from $H^2(\bar{t})$ according to eqs. 18 and 19). The prior was specified as a joint distribution of independent variables: $(g_0,\ \alpha,\ \theta,\ H^2(\bar{t}),\ \sigma_e^2) \sim \mathcal{N}(4.5, 3) \times \mathrm{Exp}(0.02) \times \mathcal{N}(4.5, 3) \times \mathcal{U}(0,\ 1) \times \mathrm{Exp}\ (0.02)$. In specifying the prior distribution, the main objective has been to use a weakly informed prior, thus, allowing the MCMC to explore a large volume of the parameter space without overwriting the signal in the data. This was verified by the nearly flat prior densities contrasting with sharply peaked posterior densities proving the presence of strong signal in the data (compare prior vs. posterior densities on supplementary fig. S8*B*, Supplementary Material online). To validate that the results were not sensitive to the parametrization and the definition of the prior, we tested other parametrizations and priors (e.g., $(\alpha,\ \theta,\ \sigma^2,\ \sigma_e^2) \sim \mathrm{Exp}(0.01) \times \mathcal{U}(0,\ 100) \times \mathrm{Exp}(0,\ 10^{-4}) \times \mathrm{Exp}(0.01)$). These resulted in matching posterior means and HPDs for all sampled and derived parameters (not reported). The adaptive Metropolis MCMC was run for 4.2E + 06 iterations, of which the first 2E + 05 were used for warm-up and adaptation of the jump distribution variance–covariance matrix. The target acceptance rate was set to 0.01 and the thinning interval was set to 1,000. The convergence and mixing of the MCMC was validated by visual analysis (supplementary fig. S8*A*, Supplementary Material online) as well as by comparison to a parallel MCMC-chain started from a different initial state. Calculation of 95% HPD was done using the function "HPDinterval" from the coda package (Plummer et al. 2006).

Computer Simulations of the Toy Epidemiological Model

The parameters defining the within- and between-host dynamics used in the simulations are written in supplementary table S2, Supplementary Material online.

The simulations were implemented as stochastic random sampling of within- and between-host events (i.e., risky contact, transmission, mutation, diagnosis, death) in discrete time-steps of length 0.05 (arbitrary time-units). The transmission history as well as the history of within-host strain substitutions was preserved during the simulations in order to reproduce exact transmission trees and to extract donor and recipient values at moments of transmission for the calculation of b_0.

Software

This study relies on two accompanying R-packages:

- toyepidemic implementing the toy epidemiological model; available at https://github.com/venelin/toyepidemic.git, last accessed January 9, 2018; and
- patherit providing a common interface for evaluating the various heritability estimators on simulated and real data. The pair correlation and regression slope estimators are implemented as functions in this package; the phylogenetic heritability estimators (PMM and POUMM) are implemented as external calls to the R-package POUMM (Mitov and Stadler 2017). The patherit package is available at https://github.com/venelin/patherit.git, last accessed January 9, 2018.

External Dependencies

The following third-party R-packages were used: ape v3.4 (Paradis et al. 2004), data.table v1.9.6 (Dowle and Srinivasan 2017), adaptMCMC v1.1 (Scheidegger 2012), Rmpfr v0.6-0 (Maechler 2016), and coda v0.18-1 (Plummer et al. 2006). All programs have been run on R v3.2.4 (R Core Team 2016).

Data Availability

All scripts for performing the simulations and real data analyses presented in this paper are available at https://github.com/venelin/Estimating-Pathogen-Trait-Heritability.git, last accessed January 9, 2018. Large output data files from the toy model simulations are available upon request to the authors. The UK HIV data are not made available at the above address, because the authors do not have the right to redistribute this data (readers are referred to the UK drug resistance database).

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

This work was supported by the Eidgenssische Technische Hochschule Zürich and in part by the European Research Council under the 7th Framework Programme of the European Commission (PhyPD: Grant Agreement Number 335529). The authors thank Dr Emma Hodcroft for sending the UK phylogeny in Newick format together with the associated spVL values, Dr Gabriel Leventhal and Prof. Sebastian Bonhoeffer for valuable insights on DR regression, Dr Francois Blanquart and Prof. Christoph Fraser for sharing with us their early results on the Beehive data set and for valuable discussions, and Dr David Rasmussen for a careful review of the manuscript.

References

Alizon S, von Wyl V, Stadler T, Kouyos RD, Yerly S, Hirschel B, Böni J, Shah C, Klimkait T, Furrer H. 2010. Phylogenetic approach reveals that virus genotype largely determines HIV set-point viral load. *PLoS Pathog.* 6(9):e1001123.

Anderson TJC, Williams JT, Nair S, Sudimack D, Barends M, Jaidee A, Price RN, Nosten F. 2010. Inferred relatedness and heritability in malaria parasites. *Proc R Soc B Biol Sci.* 277(1693):2531–2540.

Bachmann N, Turk T, Kadelka C, Marzel A, Shilaih M, Böni J, Aubert V, Klimkait T, Leventhal GE, Günthard HF, et al. 2017. Parent-offspring regression to estimate the heritability of an HIV-1 trait in a realistic setup. *Retrovirology* 14(1):33.

Bertels F, Marzel A, Leventhal G, Mitov V, Fellay J, Günthard HF, Böni J, Yerly S, Klimkait T, Aubert V, et al. 2018. Dissecting HIV virulence: heritability of setpoint viral load, CD4+ T cell decline and per-parasite pathogenicity. *Mol Biol Evol.* 35(1):27–37.

Bjorn-Mortensen K, Soborg B, Koch A, Ladefoged K, Merker M, Lillebaek T, Andersen AB, Niemann S, Kohl TA. 2016. Tracing *Mycobacterium tuberculosis* transmission by whole genome sequencing in a high incidence setting: a retrospective population-based study in East Greenland. *Sci Rep.* 6(1):33180.

Blanquart F, Wymant C, Cornelissen M, Gall A, Bakker M, Bezemer D, Hall M, Hillebregt M, Ong SH, Albert J, et al. 2017. Viral genetic variation accounts for a third of variability in HIV-1 set-point viral load in Europe. *Plos Biol.* 15(6):e2001855.

Bonhoeffer S, Fraser C, Leventhal GE. 2015. High heritability is compatible with the broad distribution of set point viral load in HIV carriers. *PLoS Pathog.* 11(2):e1004634–e1004634.

Dowle M, Srinivasan A 2017. data.table: Extension of 'data.frame'. R package version 1.10.4-3. https://CRAN.R-project.org/package=data.table.

Falconer DS, Mackay TFC. 1996. Introduction to quantitative genetics, 4th edition. Harlow, United Kingdom: Pearson Education Limited.

Felsenstein J. 1985. Phylogenies and the comparative method. *Am Nat.* 125(1):1–15.

Felsenstein J. 1988. Phylogenies and quantitative characters. *Annu Rev Ecol Syst.* 19(1):445–471.

FitzJohn RG. 2012. Diversitree: comparative phylogenetic analyses of diversification in R. *Methods Ecol Evol.* 3(6):1084–1092.

Fraser C, Hollingsworth TD, Chapman R, de Wolf F, Hanage WP. 2007. Variation in HIV-1 set-point viral load: epidemiological analysis and an evolutionary hypothesis. *Proc Natl Acad Sci U S A.* 104(44):17441–17446.

Fraser C, Lythgoe K, Leventhal GE, Shirreff G, Hollingsworth TD, Alizon S, Bonhoeffer S. 2014. Virulence and pathogenesis of HIV-1 infection: an evolutionary perspective. *Science* 343(6177):1243727–1243727.

Freckleton RP, Harvey PH, Pagel M. 2002. Phylogenetic analysis and comparative data: a test and review of evidence. *Am Nat.* 160(6):712–726.

Grimmett G, Stirzaker D. 2001. Probability and random processes. Oxford, UK: Oxford University Press.

Hansen TF. 1997. Stabilizing selection and the comparative analysis of adaptation. *Evolution Int J Org Evolution* 51(5):1341–1351.

Hansen TF, Bartoszek K. 2012. Interpreting the evolutionary regression: the interplay between observational and biological errors in phylogenetic comparative studies. *Syst Biol.* 61(3):413–425.

Hartl DL, Clark AG. 2007. Principles of population genetics. Sunderland, Massachusetts: Sinauer Associates.

Hecht FM, Hartogensis W, Bragg L, Bacchetti P, Atchison R, Grant R, Barbour J, Deeks SG. 2010. HIV RNA level in early infection is predicted by viral load in the transmission source. *AIDS* 24(7):941–945.

Hodcroft E, Hadfield JD, Fearnhill E, Phillips A, Dunn D, O'Shea S, Pillay D, Leigh Brown AJ. 2014. The contribution of viral genotype to plasma viral set-point in HIV infection. *PLoS Pathog.* 10(5):e1004112.

Hollingsworth TD, Laeyendecker O, Shirreff G, Donnelly CA, Serwadda D, Wawer MJ, Kiwanuka N, Nalugoda F, Collinson-Streng A, Ssempijja V, et al. 2010. HIV-1 transmitting couples have similar viral load set-points in Rakai, Uganda. *PLoS Pathog.* 6(5):e1000876.

Housworth EA, Martins EP, Lynch M. 2004. The phylogenetic mixed model. *Am Nat.* 163(1):84–96.

Keele BF, Giorgi EE, Salazar-Gonzalez JF, Decker JM, Pham KT, Salazar MG, Sun C, Grayson T, Wang S, Li H, et al. 2008. Identification and characterization of transmitted and early founder virus envelopes in primary HIV-1 infection. *Proc Natl Acad Sci U S A.* 105(21):7552–7557.

Keeling MJ, Rohani P. 2007. Modeling infectious diseases in humans and animals. Princeton, New Jersey: Princeton University Press.
Lande R. 1976. Natural-selection and random genetic drift in phenotypic evolution. *Evolution Int J Org Evolution* 30(2):314–334.
Leventhal GE, Bonhoeffer S. 2016. Potential pitfalls in estimating viral load heritability. *Trends Microbiol*. 24(9):687–698.
Lingappa JR, Thomas KK, Hughes JP, Baeten JM, Wald A, Farquhar C, de Bruyn G, Fife KH, Campbell MS, Kapiga S, et al. 2013. Partner characteristics predicting HIV-1 set point in sexually acquired HIV-1 among African seroconverters. *AIDS Res Hum Retroviruses*. 29(1):164–171.
Lynch M. 1991. Methods for the analysis of comparative data in evolutionary biology. *Evolution Int J Org Evolution* 45(5):1065–1080.
Lynch M, Walsh B. 1998. Genetics and analysis of quantitative traits. Sunderland, Massachusetts: Sinauer Associates Incorporated.
Lythgoe KA, Fraser C. 2012. New insights into the evolutionary rate of HIV-1 at the within-host and epidemiological levels. *Proc Biol Sci*. 279(1741):3367–3375.
Maechler M. 2016. Rmpfr: R MPFR - Multiple Precision Floating-Point Reliable. R package version 0.6-1. https://CRAN.R-project.org/package=Rmpfr
Mellors JW, Rinaldo CR, Gupta P, White RM, Todd JA, Kingsley LA. 1996. Prognosis in HIV-1 infection predicted by the quantity of virus in plasma. *Science* 272(5265):1167–1170.
Metzner KJ. 2016. HIV whole-genome sequencing now: answering still-open questions. *J Clin Microbiol*. 54(4):834–835.
Mitov V, Stadler T. 2016. The heritability of pathogen traits – definitions and estimators. bioRxiv 058503.
Mitov V, Stadler T. 2017. POUMM: An R-package for Bayesian Inference of Phylogenetic Heritability. bioRxiv 115089.
Paradis E, Claude J, Strimmer K. 2004. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* 20(2):289–290.
Plummer M, Best N, Cowles K, Vines K. 2006. CODA: convergence diagnosis and output analysis for MCMC. *R News* 6:7–11.
R Core Team. 2013. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing, https://www.R-project.org.
Scheidegger A. 2012. adaptMCMC: adaptive Monte Carlo Markov Chain sampler with coerced acceptance rate. R package version 1.1. https://CRAN.R-project.org/package=adaptMCMC.
Shirreff G, Alizon S, Cori A, Günthard HF, Laeyendecker O, van Sighem A, Bezemer D, Fraser C. 2013. How effectively can HIV phylogenies be used to measure heritability? *Evol Med Public Health* 2013(1):209–224.
Stearns SC, Koella JC. 2007. Evolution in health and disease. Oxford: OUP.
Tang J, Tang S, Lobashevsky E, Zulu I, Aldrovandi G, Allen S, Kaslow RA and Zambia-UAB HIV Research Project. 2004. HLA allele sharing and HIV type 1 viremia in seroconverting Zambians with known transmitting partners. *AIDS Res Hum Retroviruses* 20(1):19–25.
Uhlenbeck GE, Ornstein LS. 1930. On the theory of the Brownian motion. *Phys Rev*. 36(5):823–841.
van der Kuyl AC, Jurriaans S, Pollakis G, Bakker M, Cornelissen M. 2010. HIV RNA levels in transmission sources only weakly predict plasma viral load in recipients. *AIDS* 24(10):1607–1608.
Vihola M. 2012. Robust adaptive Metropolis algorithm with coerced acceptance rate. *Stat Comput*. 22(5):997–1008.
Yang Z. 2006. Computational molecular evolution. Oxford: OUP.
Yue L, Prentice HA, Farmer P, Song W, He D, Lakhi S, Goepfert P, Gilmour J, Allen S, Tang J, et al. 2013. Cumulative impact of host and viral factors on HIV-1 viral-load control during early infection. *J Virol*. 87(2):708–715.

APPENDIX

A Practical Guide to Estimating the Heritability of Pathogen Traits - SUPPLEMENTARY INFORMATION

Venelin Mitov,*[,1,2] Tanja Stadler,[1,2]
[1]Department of Biosystems, Science and Engineering (D-BSSE)
[2]Swiss Federal Institute of Technology (ETH), Zürich, Switzerland
***Corresponding author:** E-mail: vmitov@gmail.com

In the sections below, we provide additional details and evidence in support of the statements made in the main text. In section "Approximations in equations 4 and 5", we clarify the approximations used in the main text. In secton "Why does PMM underestimate the correlation between PPs in the UK-data?", we investigate in some depth the observed bad fit of the PMM model to the UK data. In section "Analysis of bias in H^2-estimates in the toy-model simulations", we explain in detail the causes of bias in H^2-estimators, which were encountered in the toy model simulations. In section "Covariance between donor and recipient values in the toy model", we show analytically that in a neutral drift scenario, when all pathogen strains are encountered at equal frequencies, the covariance between donor and recipient values decays exponentially with the evolutionary time, d_{ij} between the moments of trait measurement. In section "Choosing the threshold phylogenetic distance $d_{ij}{}'$ in ANOVA-CPP", we discuss the choice of threhold on d_{ij} (e.g. $d_{ij}{}'=D_1$ and $d_{ij}{}'=10^{-4}$) when defining closest phylogenetic pairs. Supplementary tables and figures are provided at the end of this document.

SI

Approximations in equations 4 and 5

To express the correlation in phylogenetic pairs under the PMM and the POUMM ML fits as functions of d_{ij} (eq. 4 and 5), we applied three approximations:

- In eq. 2 and 3, we replaced t by the mean root-tip distance in the tree, $\bar{t}$. This approximation was reasonable, because the mean root-tip distance did not vary substantially between different strata (fig. 3). The mean root-tip distance was 0.15 in the left-most decile going gradually down to 0.14 in the right-most decile. We also performed linear regression of the root-tip distance, t, on the phylogenetic distance, d_{ij} in the 1917 PPs. This was significant but with negligible slope and coefficient of determination ($\hat{(t)}=0.15-0.13*d_{ij}$, $p<0.01$, $R^2_{adj}=0.01$), showing that PPs of all phylogenetic distances were spread nearly uniformly across the tree. Substituting t with its linear regression on d_{ij} instead of $\bar{t}$ did not result in any noticeable difference and is not reported.

- In eq. 2 and 3, we used the relationship between t_{ij} and d_{ij}. This was the only way to incorporate d_{ij} in eq. 2. In an ultrametric tree, t_{ij} is an exact linear function of d_{ij}, namely, $t_{ij}=t-0.5d_{ij}$, where t is the root-tip distance. In the non-ultrametric UK tree, the OLS regression of t_{ij} on d_{ij} was $\hat{t_{ij}}=0.15-0.63d_{ij}$, $p<10^{-16}$, $R^2_{adj}=0.24$.
- In eq. 3, we approximated $\exp(-8.35+36.47d_{ij})$ with 0, which was a valid approximation on the scale of the other terms in the equation and for the range of phylogenetic distances ($d_{ij}\in[0,0.14]$) in the UK tree.

The above approximates were validated visually by comparing the analytical curves corresponding to equations 4 and 5 with the corresponding brown and green points and error-bars on fig. 3.

Why does PMM underestimate the correlation between PPs in the UK data?

We have shown in the main text that, the phenotypic correlation between members of phylogenetic pairs depends on their phylogenetic distance, d_{ij}: members of pairs with small d_{ij} tend to have higher phenotypic correlation compared to members of pairs with big d_{ij} (fig. 3). For PMM, the only way to incorporate this information is indirect, namely, through the relationship between d_{ij} and the root-mrca distance, t_{ij}. In the non-ultrametric UK tree, this relationship is rather weak: the slope of the OLS regression of d_{ij} on t_{ij} equals -0.37 and is significant ($p<0.01$) but the coefficient of determination of this regression, R^2_{adj}, is (only) 0.24 (fig. S2A). Thus, the principal source of information for fitting the PMM parameters, σ^2 and σ^2_e, is the assumed linear relationship between the observable correlation between pairs of tips and the two distances involved in eq. 2: the root-mrca distance t_{ij} and the root-tip distance, t. Noticing that the correlation between the lg(spVL)-values in phylogenetic pairs is a covariance to variance ratio (eq. 2 and 3), we analyze how PMM fits to these two components in the UK data (fig. S2 B and C). The panels B and C on fig. S2 show that the covariance and the variance progress at different rates with t_{ij} and t respectively. PMM is not able to model this difference in the rates, because it uses a single parameter, σ^2, to model both of them. We notice that the ML estimate for σ^2 fits well to the linear increase in the variance (parallel brown and black lines on fig. S2C) but underestimates the increase in the covariance (non-parallel brown and black lines on fig. S2B). This indicates that a linear model of the covariance as a function of t_{ij} is rather inappropriate and no particular value for σ^2 could result in a better fit (higher likelihood). As a result, the penalty on the PMM likelihood is minimized when the parameter σ^2 is fit to the increase in the variance, neglecting the covariance. Finally, this leads to the observed underestimate of the correlation in the closest phylogenetic pairs.

Analysis of bias in H^2-estimates in the toy-model simulations

In order to understand the origin of the bias in the different toy-model scenarios, we used variance decomposition into the heritable component, σ_G^2 and the non-heritable component σ_e^2. Most of the biases observed on fig. 4 could be explained by a bias in one or both of these two components. The main source of these biases was the within-host evolution causing a decrease in the measured covariance between donor-recipient partners or phylogenetic pairs. Also, we identified various sampling biases introduced by within-/between-host selection and filtering of the data. We clarify these sources of bias in the following subsections.

Neutral evolution of the trait within hosts

This phenomenon consists in a random change of the trait value caused by pathogen mutation. As a result, the phenotypic correlation between donors and recipients tends to decrease. We show later that, in a neutral scenario, this correlation decay is expected to be exponential in the phylogenetic distance, d_{ij}. As a result, all H^2-estimators neglecting or improperly modeling this decay are negatively biased. The most affected estimators are $b_{d_{ij}}$, $r_{A,d_{ij}}$, $H^2_{BM}(\bar{t})$ and H^2_{BMe} (fig. 4); see also the decreasing sample donor-recipient covariance $s(z_{don}, z_{rcp})$ on fig. S4.

Directional selection within a host

This phenomenon consists in mutant strains contributing to a higher trait value, e.g. strains with higher reproductive capacity in the case of viral load, getting selected within each host. As a result a population of newly infected hosts tends to have higher genotypic variance than a population of hosts which have undergone within-host evolution (notice $s^2(G_{rcp,0}) > s^2(G)$ on fig. S4B). This explains the positive bias of b_0 with respect to H^2 on fig. 4B in the main text. Another possible effect of within-host selection is a convergent evolution in donors and recipients towards strains, which have higher fitness on average in the population. Intuitively, this could lead to a slight increase in phenotypic covariance and, therefore, a positive bias in b. Such a bias was not obvious in the toy-model simulations ($s(z_{don,0}, z_{rcp,0}) > s(z_{don}, z_{rcp})$ in all simulations, fig. S4B), because the convergent evolution was leading to a decreasing overall genetic and phenotypic variance in the population (see decreasing $s^2(G)$ and $s^2(z)$ with d_{ij} on fig. S4B and S5B).

Stabilizing selection between hosts

In case the trait is positively correlated with pathogen load, virulence and per contact transmission rate, hosts with very high pathogen load tend to be more infectious but stay infectious for a short period of time due to earlier diagnosis or death; hosts with very low pathogen load are infectious for a longer time but transmit very rarely. Thus, hosts with intermediate values of pathogen load have the highest

transmission potential on average (Fraser et al. 2007). This leads to a sampling bias in donor-recipient estimators - the donors have a narrower distribution than the overall population ($s^2(z_{don}) < s^2(z)$ on Fig. S5C). Intuitively this should lead to a positive bias in b_0 with respect to H^2 (because the denominator ($s^2(z_{don})$ is smaller). However, this was not confirmed by the toy-model simulations because the genotypic variance in the donor-recipient values at the moment of transmission was also smaller than that at the population level ($s(z_{don,0}, z_{rcp,0}) \approx s^2(G_0) < s^2(G)$ on fig. S4C).

Combined within- and between-host selection

This results in a combination of the sampling biases due to each of the two selection phenomena (previous subsections).

Non-stationary trait distribution during the epidemic

The density of the trait values evolves during the epidemic due to continuous change in the frequencies of the different pathogen strains, introduction of new strains through de-novo mutation, change in the frequencies of infected host types and a number of demographic factors such as migration, prevention, diagnosis and treatment. Thus, the broad-sense heritability, H^2, is a dynamic property of the population which changes through time. The direct estimator R^2_{adj} obtained over a grouping by identical strain in patients sampled at different times has the meaning of a summary statistic averaging over the time of the epidemic. Plotting the phylogenetic estimators $H^2_{BM}(\sigma,\sigma_e,t)$ and $H^2_{OU}(\alpha,\sigma,\sigma_e,t)$ over time can help understanding the above dynamics. This, however, depends strongly on the goodness of fit of the phylogenetic model (e.g. BM or OU) to the data.

Violation of phylogenetic model assumptions

The phylogenetic estimates of heritability are valid only if the model assumptions are at least partially met. For example, in this article, we have shown how an inaccurate assumption about the form of the correlation between two tips in the PMM model can lead to a significant negative bias in phylogenetic heritability.

Clarifying the observed positive bias in r_A[id]

Here we demonstrate a positive bias in r_A with respect to R^2_{adj} for small number of groups (genotypes) K. We show that this bias vanishes for bigger values of K, i.e. $K > 24$, given that the genotypic values are sampled from a normal distribution. For each $K \in \{3,6,12,24,48\}$ we simulate 100 datasets with K genotypes and varying number of carriers for each genotype. We draw genotypic values from a normal distribution and add random (white) noise to them to construct the phenotype. After estimating R^2, R^2_{adj} and r_A for each dataset, we report the average values for each K.

```
library(data.table)
library(patherit)
# grand mean and variance of group effects
mu <- 3.5
sigma2a <- .2
# within-class variance
sigma2e <- 0.36

#number of simulated data-sets with K groups and ni individuals per group
nIter <- 100

# make results reproducible
set.seed(20)

test <- list()

# number of classes/groups
for(K in c(3, 6, 12, 24, 48, 96)) {

  test[[as.character(K)]] <- t(sapply(1:nIter, function(iter) {
    # sample group means at each iteration from a normal distribution
    ai <- rnorm(K, mean=mu, sd=sqrt(sigma2a))
    # numbers of sampled individuals per group
    ni <- sample(20:50, K, replace=TRUE)
    # generate data
    data <- data.table(g=do.call(c, lapply(1:K, function(k) rep(k, ni[k]))), key='g')
    data[, z:=rnorm(ni[g], mean=ai[g], sd=sqrt(sigma2e)), by=g]
    data[, G:=mean(z), by=g]
    data[, e:=z-G]
    rAValues <- rA(epidemic=NULL, data=data, GEValues=NULL, by='g', report=TRUE)
    with(rAValues, data[, c(K=K, H2true=sigma2a/(sigma2a+sigma2e),
                            R2=var(G)/var(z), R2adj=1-(N-1)/(N-K)*var(z-G)/var(z),
                            rA=H2aov)])
  }))
}

t(sapply(names(test), function(K) {
  colMeans(test[[K]])
}))
```

```
##      K    H2true        R2     R2adj        rA
## 3    3 0.3571429 0.2304160 0.2147992 0.2768490
## 6    6 0.3571429 0.2987824 0.2816688 0.3181378
## 12  12 0.3571429 0.3475622 0.3298251 0.3494137
## 24  24 0.3571429 0.3622040 0.3441696 0.3539454
## 48  48 0.3571429 0.3655049 0.3472988 0.3522200
## 96  96 0.3571429 0.3720558 0.3537375 0.3562131
```

The results show that r_A dominates R^2_{adj} on average, in particular for small values of K, i.e. $K \leqslant 12$. For bigger K, the two estimators are asymptotically equal.

Clarifying the observed difference between $H^2_{BM}(\bar{t})$ and H^2_{BMe} in toy-model simulations

As another detail, we notice that the expected correlation under the PMM ML fit, r_{BM}, was significantly positively biased with respect to the correlation, r_A, measured in PPs (brown line versus brown dots on fig. S7). Investigating these cases, we found that these positive biases were due to the use of the mean root-tip distance $\bar{t}$ in the formulation of r_{BM} (eq. 2), the bias being less pronounced if using the median or a higher quantile of the root-tip distance. Since, at $d_{ij}=0$, r_{BM} is equal to the phylogenetic heritability, $H^2_{BM}(\bar{t})$, in these cases, we observe a value of $H^2_{BM}(\bar{t})$ closer to the true heritability value, R^2_{adj} (black horizontal line on fig. S7). This could lead to a wrong conclusion that $H^2_{BM}(\bar{t})$ is less biased than H^2_{BMe}. In fact though, this is merely the effect of cancelling out two biases with opposite directions. Compared to the PMM, the POUMM produced a better fit to the decaying correlation in all simulations (green dots and error-bars on fig. S7).

Covariance between donor and recipient values in the toy model

One of the main results of this article is the observation that the accuracy of a heritability estimator depends strongly on how it accounts for the within-host evolution of the pathogen taking place between transmission events and measurement. This becomes obvious from the fact that both, real data and the toy model simulations, showed a pattern of decaying correlation between phylogenetic pairs as a function of their phylogenetic distance, d_{ij} (fig. 3 and fig. S7). Is this pattern of decaying correlation a general characteristic of epidemics? Here, we use a simplified version of the toy model allowing an analytical approach to this question.

We consider a version of the toy model, in which there is one SNP in the pathogen genotype with two possible alleles and there are two possible host-types. We denote the four genotype×host-type combinations as subscripts 00, 01, 10, 11, where the first index denotes the pathogen genotype and the second index denotes the host-type. The trait values are denoted as z_{00}, z_{01}, z_{10} and z_{11}; the frequencies of the four genotype×host-type combinations in the populations are denoted as f_{00}, f_{01}, f_{10} and f_{11}. We use the symbol $f_{\cdot 0}=f_{00}+f_{10}$ to denote the total frequency of host-type 0 and $f_{\cdot 1}=f_{01}+f_{11}$ to denote the total frequency of host-type 1. We assume that the evolution of the pathogen strain within a host follows random drift - at time $d_{ij}/2$ after infection, the strain infecting a host has been substituted by a mutant strain with probability ν, regardless of the trait value before and after substition; the strain has remained unchanged with probability $1-\nu$.

This mechanism of within-host mutation is summarized on fig. S9A. For simplicity, we assume a generation-based dynamics, in which transmission to new susceptible hosts occurs at fixed moments in

time separated by a period $d_{ij}/2$. At every generation, each member of the infected population transmits his/her currently carried pathogen to a random susceptible individual and becomes uninfectious, (although, he/she remains infected with the pathogen). The recipient host transmits his infection at the next generation and becomes uninfectious on his turn. We assume an infinite susceptible pool with fixed frequencies of the two host-types. Given that there is no selection with respect to host-type, we can assume that the frequencies of the host-types in the infected population equals the host-type frequencies in the susceptible population. The frequencies of the two pathogen strains in the infected population can evolve as a result of within-host mutation. However, in the absence of within-host selection, the strain frequencies conditioned on host-type would equalize several generations after the onset of epidemic.

With the above simplified version of the toy model, it is possible to express the covariance between a donor and recipient trait value at time $d_{ij}/2$ after the transmission has taken place. We do this in two steps: first, we express the covariance in terms of the substitution probability ν; then, we use a 2-nucleotide form of the Jukes-Cantor 69 substitution model to express ν in terms of evolutionary time. Denoting the donor value by z_{don} and the recipient value by z_{rcp}, we start from a known property for the covariance:

$$Cov(z_{don}, z_{rcp}) = E[z_{don} z_{rcp}] - E[z_{don}]E[z_{rcp}]$$

In the case of neutral evolution and sufficiently large population size, we can assume that donors and recipients share the same frequencies of genotype $\times$ host-type combinations. Thus, we write:

$$E[z_{don}] = E[z_{rcp}] = f_{00} z_{00} + f_{01} z_{01} + f_{10} z_{10} + f_{11} z_{11}$$

To obtain the expectation of the product $z_{don} z_{rcp}$, it suffices to sum up all possible products of donor and recipient values weighted by their expected frequencies (fig. S9A):

$$
\begin{aligned}
E[z_{don}z_{rcp}] = f_{00}\Big(&(1-\nu)f_{\cdot 0}(1-\nu)z_{00}z_{00}+(1-\nu)f_{\cdot 0}\nu z_{00}z_{10}+(1-\nu)f_{\cdot 1}(1-\nu)z_{00}z_{01}+(1-\nu)f_{\cdot 1}\nu z_{00}z_{11}+\\
&\nu f_{\cdot 0}(1-\nu)z_{10}z_{00}+\nu f_{\cdot 0}\nu z_{10}z_{10}+\nu f_{\cdot 1}(1-\nu)z_{10}z_{01}+\nu f_{\cdot 1}\nu z_{10}z_{11}\Big)+\\
f_{01}\Big(&(1-\nu)f_{\cdot 0}(1-\nu)z_{01}z_{00}+(1-\nu)f_{\cdot 0}\nu z_{01}z_{10}+(1-\nu)f_{\cdot 1}(1-\nu)z_{01}z_{01}+(1-\nu)f_{\cdot 1}\nu z_{01}z_{11}+\\
&\nu f_{\cdot 0}(1-\nu)z_{11}z_{00}+\nu f_{\cdot 0}\nu z_{11}z_{10}+\nu f_{\cdot 1}(1-\nu)z_{11}z_{01}+\nu f_{\cdot 1}\nu z_{11}z_{11}\Big)+\\
f_{10}\Big(&(1-\nu)f_{\cdot 0}(1-\nu)z_{10}z_{10}+(1-\nu)f_{\cdot 0}\nu z_{10}z_{00}+(1-\nu)f_{\cdot 1}(1-\nu)z_{10}z_{11}+(1-\nu)f_{\cdot 1}\nu z_{10}z_{01}+\\
&\nu f_{\cdot 0}(1-\nu)z_{00}z_{10}+\nu f_{\cdot 0}\nu z_{00}z_{00}+\nu f_{\cdot 1}(1-\nu)z_{00}z_{11}+\nu f_{\cdot 1}\nu z_{00}z_{01}\Big)+\\
f_{11}\Big(&(1-\nu)f_{\cdot 0}(1-\nu)z_{11}z_{10}+(1-\nu)f_{\cdot 0}\nu z_{11}z_{00}+(1-\nu)f_{\cdot 1}(1-\nu)z_{11}z_{11}+(1-\nu)f_{\cdot 1}\nu z_{11}z_{01}+\\
&\nu f_{\cdot 0}(1-\nu)z_{01}z_{10}+\nu f_{\cdot 0}\nu z_{01}z_{00}+\nu f_{\cdot 1}(1-\nu)z_{01}z_{11}+\nu f_{\cdot 1}\nu z_{01}z_{01}\Big)
\end{aligned}
$$

Taking the difference of $E[z_{don}z_{rcp}]-E[z_{don}]E[z_{rcp}]$ and grouping on the degrees of ν, we obtain a polynomial of degree two of ν:

$$Cov(z_{don},z_{rcp}) = A\nu^2+B\nu+C$$

The coefficients A, B and C are algebraic expressions of the frequencies and trait values:

$$
\begin{aligned}
A &= (f_{00}(z_{00}-z_{10})+f_{10}(z_{00}-z_{10})+(f_{01}+f_{11})(z_{01}-z_{11}))(f_{\cdot 0}(z_{00}-z_{10})+f_{\cdot 1}(z_{01}-z_{11}))\\
B &= 2f_{10}f_{\cdot 0}z_{00}z_{10}+f_{11}f_{\cdot 0}z_{01}z_{10}+f_{10}f_{\cdot 1}z_{01}z_{10}-2f_{10}f_{\cdot 0}z_{10}^2+f_{11}f_{\cdot 0}z_{00}z_{11}+\\
&\quad f_{10}f_{\cdot 1}z_{00}z_{11}+2f_{11}f_{\cdot 1}z_{01}z_{11}-2f_{11}f_{\cdot 0}z_{10}z_{11}-2f_{10}f_{\cdot 1}z_{10}z_{11}-2f_{11}f_{\cdot 1}z_{11}^2+\\
&\quad f_{01}(2f_{\cdot 1}z_{01}(-z_{01}+z_{11})+f_{\cdot 0}(-2z_{00}z_{01}+z_{01}z_{10}+z_{00}z_{11}))+\\
&\quad f_{00}(-2f_{\cdot 0}z_{00}(z_{00}-z_{10})+f_{\cdot 1}(-2z_{00}z_{01}+z_{01}z_{10}+z_{00}z_{11}))\\
C &= f_{00}f_{\cdot 0}z_{00}^2+f_{01}f_{\cdot 0}z_{00}z_{01}+f_{00}f_{\cdot 1}z_{00}z_{01}+f_{01}f_{\cdot 1}z_{01}^2+f_{10}f_{\cdot 0}z_{10}^2+f_{11}f_{\cdot 0}z_{10}z_{11}+\\
&\quad f_{10}f_{\cdot 1}z_{10}z_{11}+f_{11}f_{\cdot 1}z_{11}^2-(f_{00}z_{00}+f_{01}z_{01}+f_{10}z_{10}+f_{11}z_{11})^2
\end{aligned}
$$

In the case of neutral drift, $f_{00}=f_{10}$ and $f_{01}=f_{11}$. Substituting $1-f_{\cdot 0}$ for $f_{\cdot 0}$, the expression for the covariance simplifies to:

$$Cov(z_{don},z_{rcp}) = \frac{1}{4}(1-2\nu)^2(z_{01}-z_{11}+f_{\cdot 0}(z_{00}-z_{01}-z_{10}+z_{11}))^2$$

Assuming a two-nucleotide Jukes Cantor 69 model with mutation rate λ, the probability of mutation at a site in the genetic sequence is expressed as a function of evolutionary time $\nu(t)=0.5-0.5\exp(-\lambda t)$ (Yang, 2006). Thus, in the case of neutral drift, the covariance between the donor and the recipient

value at time $d_{ij}/2$ after the transmission can be expressed in terms of the total evolutionary time, d_{ij}, separating the two hosts:

$$Cov(z_{don}, z_{rcp}) = \exp(-\lambda d_{ij}) \frac{1}{4} (z_{01} - z_{11} + f_{\cdot 0}(z_{00} - z_{01} - z_{10} + z_{11}))^2$$

The above expression for the covariance between donor and recipient values represents an exponential decay function of d_{ij} - it has a non-negative value at $d_{ij} = 0$ and converges exponentially towards 0 as $d_{ij} \to \infty$. It is interesting to ask wether the above pattern of exponentially decaying covariance is preserved in the case of multiple loci (many possible pathogen genotypes) as well as in cases of within- and between-host selection. An analytical treatment of this question is beyound the scope of this article. However, using simulations of the toy model, we have shown that the pattern of exponential decay seems to be preserved in the case of the neutral/neutral scenario, that is, when each pathogen genotype is encountered at equal frequency for each host-type (fig. S7). Biologically, this reflects a situation, where the donor and recipient host exhibit similar trait values shortly after transmission, but later on tend to have uncorrelated values as a result of random mutation in the two hosts. In infinite time after the transmission, the correlation between the two hosts' trait values should converge to 0. In the cases of within- or between-host selection, the covariance between the trait values of a donor and a recipient would be influenced by additional factors such as similar age, race or habitat. This can result in convergent evolution of the pathogens within the two hosts towards strains which are best adapted to the shared environmental conditions. In this case, the covariance would deviate substantially from an exponential decay function of d_{ij} and is even not guaranteed to converge to 0. This reaffirms that any parametric model of the covariance (and therefore, correlation) between transmission couples needs to be validated against empirical estimates (see fig. 3 and Fig. S2).

Choosing the threshold phylogenetic distance $d_{ij}{}'$ in ANOVA-CPP

Choosing an appropriate value for the threshold $d_{ij}{}'$ is one of the tricky aspects of ANOVA-CPP. This choice is a trade-off between minimizing the negative bias due to within-host evolution ($d_{ij}{}'$ close to 0) and maximizing the precision in terms of narrow confidence interval. While it is impossible to measure the bias in the absence of knowledge about the true value, there are ways to measure the precision, e.g. by taking the length of the 95% confidence interval. Thus, one way to define an optimality criterion is "the minimum value of $d_{ij}{}'$, for which the 95% confidence interval is narrower than some predefined length". A practical way to do this is to consider different stratifications of the phylogenetic pairs as shown on fig. S1. In the toy-model simulations, we have chosen the first decile, i.e. $d_{ij}{}' = D_1$, because this threshold

was suitable for demonstrating the negative bias due to within-host evolution (i.e. a difference between b_{D_1} and $b_{d_{ij}}$ and loss of precision). In the real HIV data, the choice $d_{ij}{}'=10^{-4}$ was based on empirical observations (see text and fig. 5).

SUPPLEMENTARY TABLES

Table S1. PMM and POUMM fit to lg(spVL) data from the UK HIV cohort.

N	Model	AICc	Type	g_0	α	θ	σ	σ_e	$H^2(\bar{t})$	H_e^2
8,483	PMM	21,487	MLE	4.49	-	-	0.65	0.84	0.08	0.06
			Mean	4.49	-	-	0.67	0.83	0.08	0.06
			HPD	[4.31, 4.66]	-	-	[0.5, 0.84]	[0.82, 0.85]	[0.05, 0.12]	[0.02, 0.1]
	POUMM	21,455	MLE	5.54	28.78	4.45	2.97	0.77	0.21	0.2
			Mean	5.44	-	-	3.11	0.77	0.21	0.21
			HPD	[4.06, 7.25]	[16.64, 46.93]	[4.41, 4.49]	[1.95, 4.37]	[0.73, 0.8]	[0.14, 0.29]	[0.13, 0.29]

Table S2. Within- and between-host dynamics of the toy epidemiological model.

Scope	Parameter	neutral	select
Between-host	Natural birth rate	$\lambda = 117.6$	
	Natural per capita death rate	$\mu = 1/850$	
	Per capita recovery rate	$\rho = 1/48$	
	Per capita contact rate	$\kappa \in \{\frac{1}{2}, \frac{1}{4}, \frac{1}{6}, \frac{1}{8}, \frac{1}{10}, \frac{1}{12}\}$	
	Per capita risky contact rate (S: current proportion of susceptible in the pop.)	$S \times \kappa$	
	Per risky contact transmission probability	$\gamma_{\text{neutral}} = .45$	$\gamma(z) = \gamma_{\min} + \frac{(\gamma_{\max} - \gamma_{\min})(\gamma_{50})^{\gamma_k}}{10^{z\gamma_k} + (\gamma_{50})^{\gamma_k}}$, where $\gamma_{\min} = .3, \gamma_{\max} = .6, \gamma_{50} = 10^3, \gamma_k = 1.4$
	Per capita death rate for infected individuals	$\delta_{\text{neutral}} = .01$	$\delta(z) = \mu + \frac{10^{zD_k} + (D_{50})^{D_k}}{D_{\min} 10^{zD_k} + D_{\max}(D_{50})^{D_k}}$, where $D_{\min} = 2, D_{\max} = 300, D_{50} = 10^3, D_k = 1.4$
Within-host	Per locus pathogen mutation rate	$\nu_{\text{neutral}} = .01$	$\nu(z) = \frac{\nu_{\max}(\nu_{50})10^{z\nu_k}}{10^{z\nu_k} + (\nu_{50})^{\nu_k}}$, where $\nu_{\max} = .2, \nu_{50} = 10^3, \nu_k = 1.4$
	Rate of substitution of strain $\mathbf{x}_j$ for $\mathbf{x}_i$, where $\mathbf{x}_i \neq \mathbf{x}_j$ at a single locus, l, M_l is the number of alleles at locus l, and the corresponding values are z_i and z_j	$\xi_l = \frac{\nu_{\text{neutral}}}{M_l - 1}$	$\xi_{l,i \leftarrow j}(z_i, z_j) = \begin{cases} \frac{\nu(z_i)}{M_l - 1} \text{ if } \nu(z_i) < \nu(z_j) \\ 0 \quad \text{, otherwise} \end{cases}$

SUPPLEMENTARY FIGURES

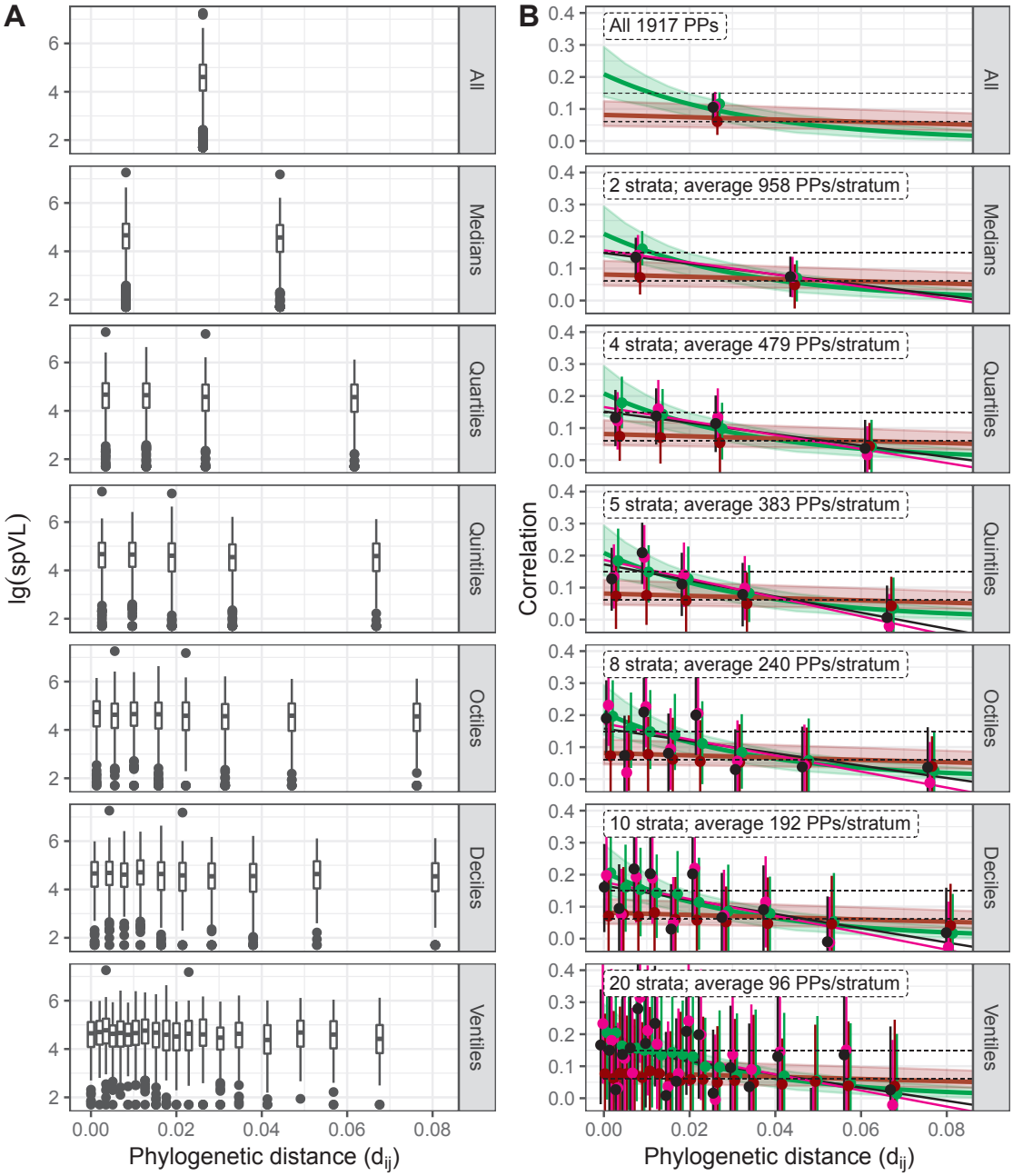


FIG. S1. Different stratifications of the phylogenetic pairs in the UK tree. A - box-plots of the trait values show nearly identical distributions (equal mean and interquartile range) in the different strata. B - correlation profiles in different stratifications. Black and magenta points with error-bars denote the estimated r_A and r_{Sp} in the real data. Dashed horizontal bars denote the 95% CI for r_A evaluated on all phylogenetic pairs. A black and a magenta inclined line denote the least squares linear regression of r_A and r_{Sp} on the mean phylogenetic distance, $\bar{d_{ij}}$, in each decile. Brown and green points with error bars denote the estimated values of r_A obtained after replacing the real trait values on the tree by values simulated under the maximum likelihood fit of the PMM and the POUMM methods respectively (mean and 95% CI estimated from 100 replications). A brown and a green line show the expected correlation between pairs of tips at distance d_{ij}, as modeled under the ML-fit of the PMM and the POUMM (eq. 2 and 3). A light-brown and a light-green region depict the 95% high posterior density (HPD) intervals inferred from Bayesian fit of the two models ("Materials and methods").

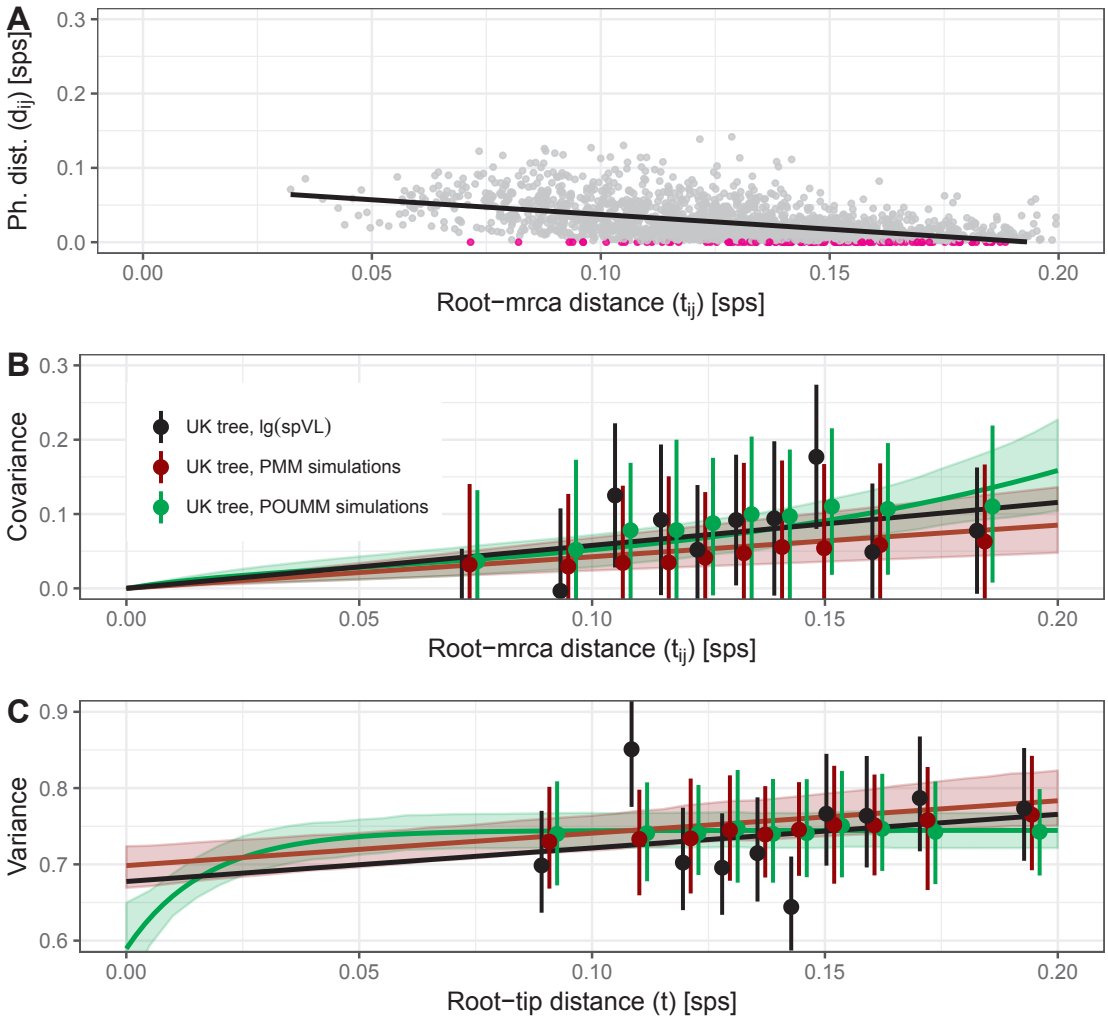


FIG. S2. Bias in the PMM estimate for the correlation in phylogenetic pairs A: A scatter plot and OLS regression of d_{ij} on t_{ij} (slope -0.34 ($p<0.01$), $R^2_{adj}=0.24$). Points in magenta denote CPPs ($d_{ij}<10^{-4}$); B: covariance modeled as a function of t_{ij}. Black points and error-bars denote the sample covariance and 95% CI upon a stratification in deciles of t_{ij}. Brown and green points and error-bars denote the mean and 95% CI upon replacing the lg(spVL)-values with values simulated under the ML fit of the PMM and the POUMM (100 replications). A black line going through the origin denotes the OLS regression with 0 intercept to the real data. A brown and a green line with brighter surrounding regions denote the covariance and its 95% HPDs under the PMM and the POUMM respectively. The latter have been obtained from the expressions for the nominator in eqs. 2 and 3 using the ML estimates and posterior samples for the model parameters. In the case of the POUMM, the phylogenetic distance d_{ij} has been replaced by the linear regression of d_{ij} on t_{ij} from the phylogenetic pairs (panel A). The slope of the brown line equals the parameter σ^2 of the PMM. Notice the negative bias with respect to the OLS fit (black line). C: variance of the trait values at the tips of the UK tree modeled as a function of the root-tip distance, t. Black points and error-bars denote the sample variance and its 95% CI in the real data, upon a stratification in deciles. Brown and green points and error-bars denote the mean and 95% CI upon replacing the lg(spVL)-values with values simulated under the ML fit of the PMM and the POUMM (100 replications). A black line denotes the OLS regression of the variance in the real data on t. A brown and a green line with brighter surrounding regions denote the variance and its 95% HPDs under the PMM and the POUMM respectively. The latter have been obtained from the expressions for the denominator in eqs. 2 and 3 using the ML estimates and posterior samples. As in panel B, the slope of the brown line equals the parameter σ^2 of the PMM. The distances t_{ij} and d_{ij} are measured in substitutions per site (sps).

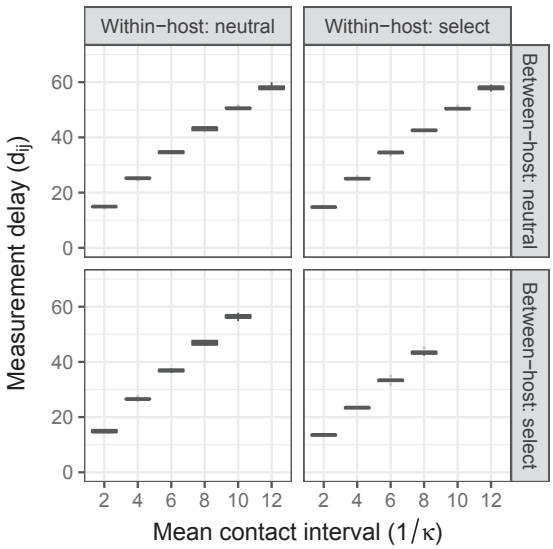


FIG. S3. Mean phylogenetic distance d_{ij} between PPs in the toy-model simulations

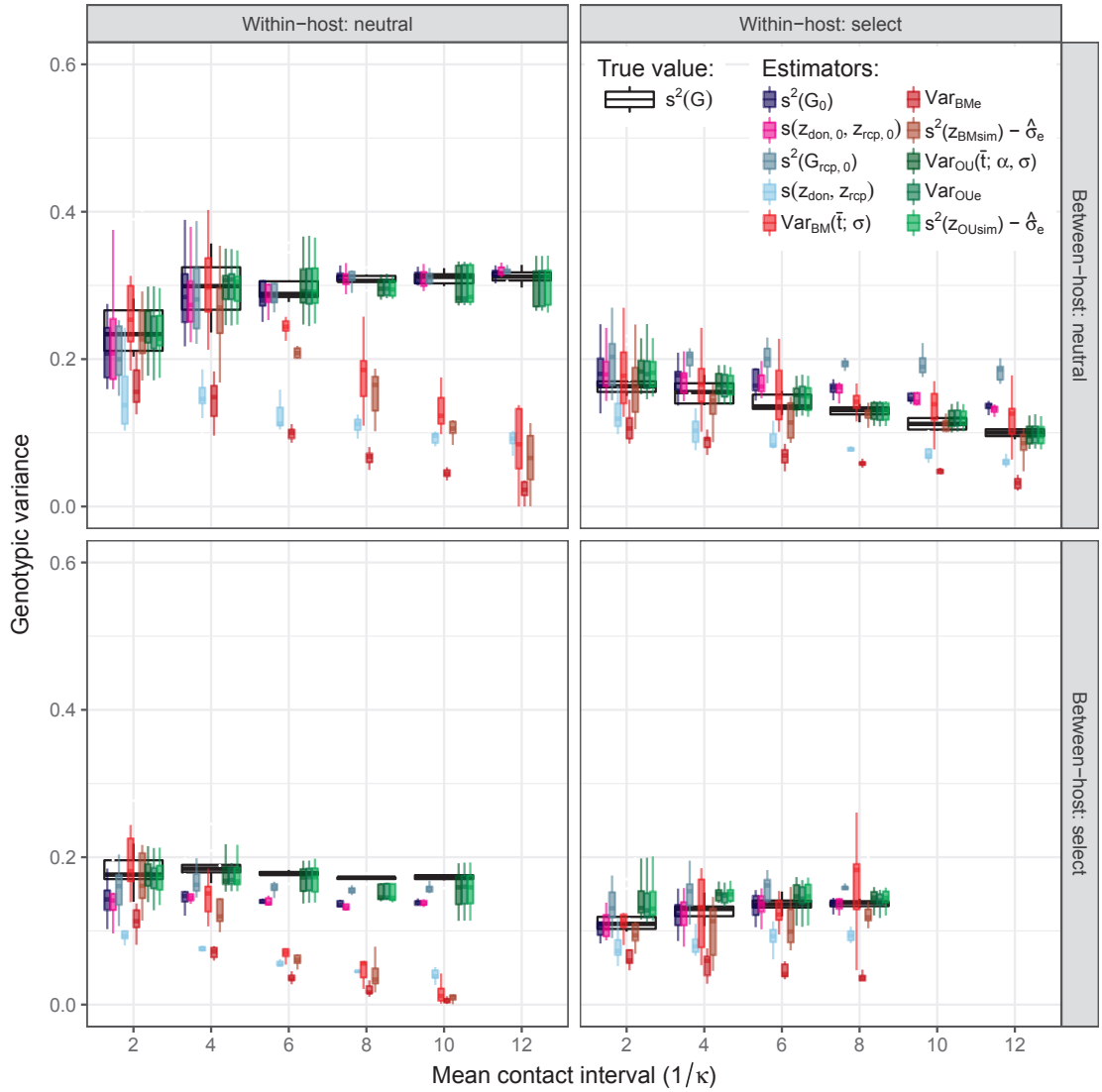


FIG. S4. Estimating the genotypic variance in toy-model simulations. $s^2(G)$: true genotypic variance calculated from grouping by identical genotype; $s^2(G_0)$: true genotypic variance calculated from grouping by identical genotype in the sample of known donor-recipients, taking their genotype and trait values at the moment of transmission; $s(z_{don,0},z_{rcp,0})$: empirical covariance between donors and recipients at the moment of transmission; $s^2(G_{rcp,0})$: true genotypic variance in recipients at the moment of getting infected; $s(z_{don},z_{rcp})$: donor-recipient covariance at moment of diagnosis (including measurement delay); $\mathrm{Var}_{BM}(\bar{t};\sigma)$: estimated PMM genotypic variance at $\bar{t}$ according to eq. 16; Var_{BMe}: estimated PMM genotypic variance based on the difference $s^2(z)-\hat{\sigma}_e^2$ in the ML fit of the PMM; $s^2(z_{BMsim})-\hat{\sigma}_e^2$: estimated PMM genotypic variance based on the difference of the mean trait variance in 100 simulations of the ML PMM fit on the tree and the ML value of the parameter σ_e^2; $\mathrm{Var}_{OU}(\bar{t};\alpha,\sigma)$: estimated POUMM genotypic variance at $\bar{t}$ according to eq. 17; Var_{OUe}: estimated POUMM genotypic variance based on the difference $s^2(z)-\hat{\sigma}_e^2$ in the ML fit of the POUMM; $s^2(z_{OUsim})-\hat{\sigma}_e^2$: estimated POUMM genotypic variance based on the difference of the mean trait variance in 100 simulations of the ML POUMM fit on the tree and the ML value of the parameter σ_e^2.

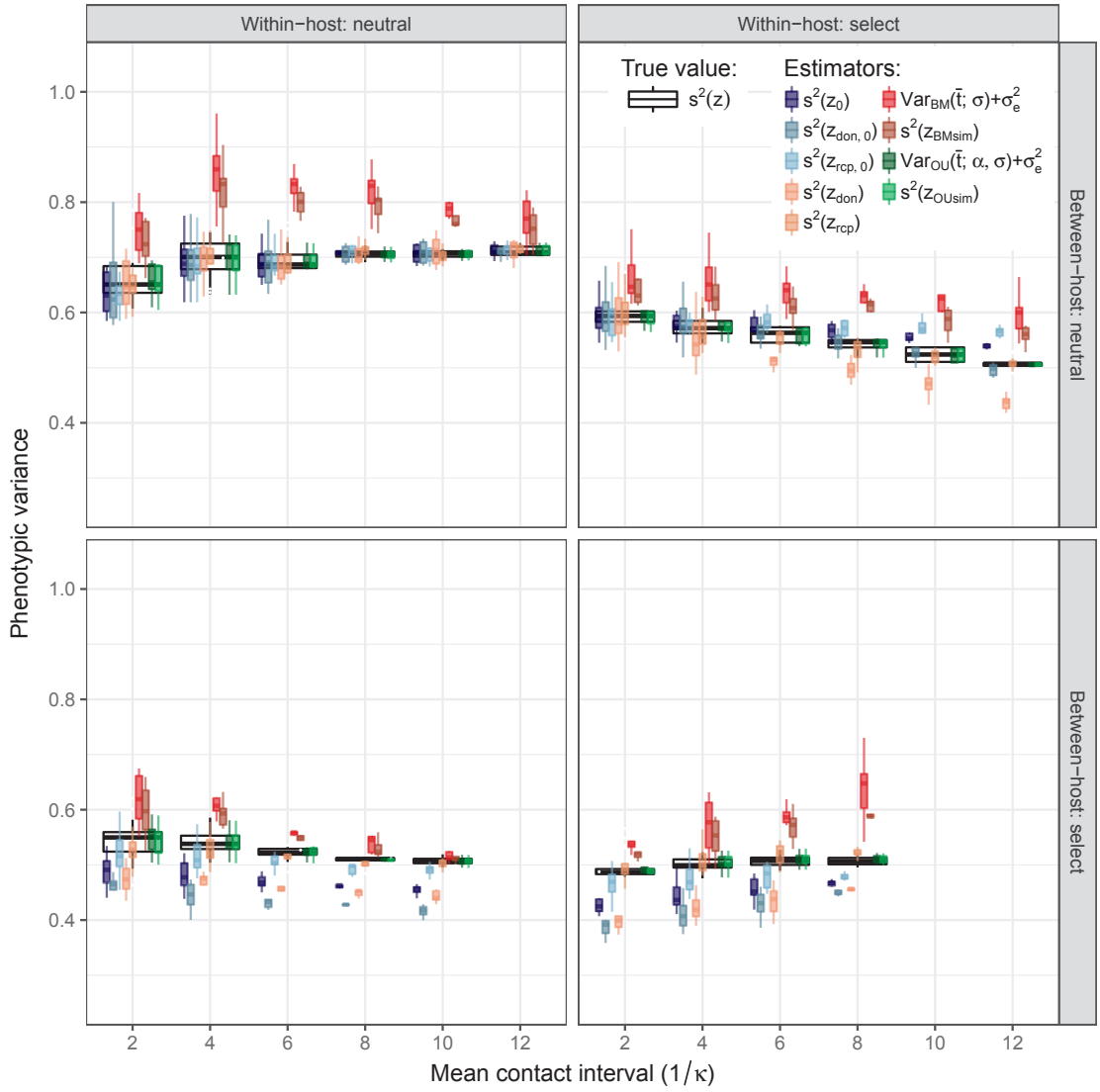


FIG. S5. Phenotypic variance in the toy-model simulations. $s^2(z)$: sample variance of the trait value in the entire sampled population; $s^2(z_0)$: sample variance in the sampled donor-recipient couples taking the trait values at moment of infection; $s^2(z_{don,0})$: sample variance in the donors from donor-recipient couples, taking the trait values at moment of infection; $s^2(z_{rcp,0})$: sample variance in the recipients from donor-recipient couples, taking the trait values at moment of infection; $s^2(z_{don})$: sample variance in the donors from donor-recipient couples, taking the trait values at moment of diagnosis; $s^2(z_{rcp})$: sample variance in the recipients from donor-recipient couples, taking the trait values at moment of diagnosis; $\mathrm{Var}_{BM}(\bar{t};\sigma,\sigma_e)=\sigma^2\bar{t}+\sigma_e^2$: expected phenotypic variance under the ML fit of the PMM at the mean root-tip distance $\bar{t}$; $s^2(z_{BMsim})$: mean sample trait variance from 100 simulations of the ML PMM fit on the tree; $\mathrm{Var}_{OU}(\bar{t};\alpha,\sigma,\sigma_e)=\frac{\sigma^2}{2\alpha}(1-\exp(-2\alpha t)+\sigma_e^2$: expected phenotypic variance under the ML fit of the POUMM at the mean root-tip distance $\bar{t}$; $s^2(z_{OUsim})$: mean sample trait variance from 100 simulations of the ML POUMM fit on the tree.

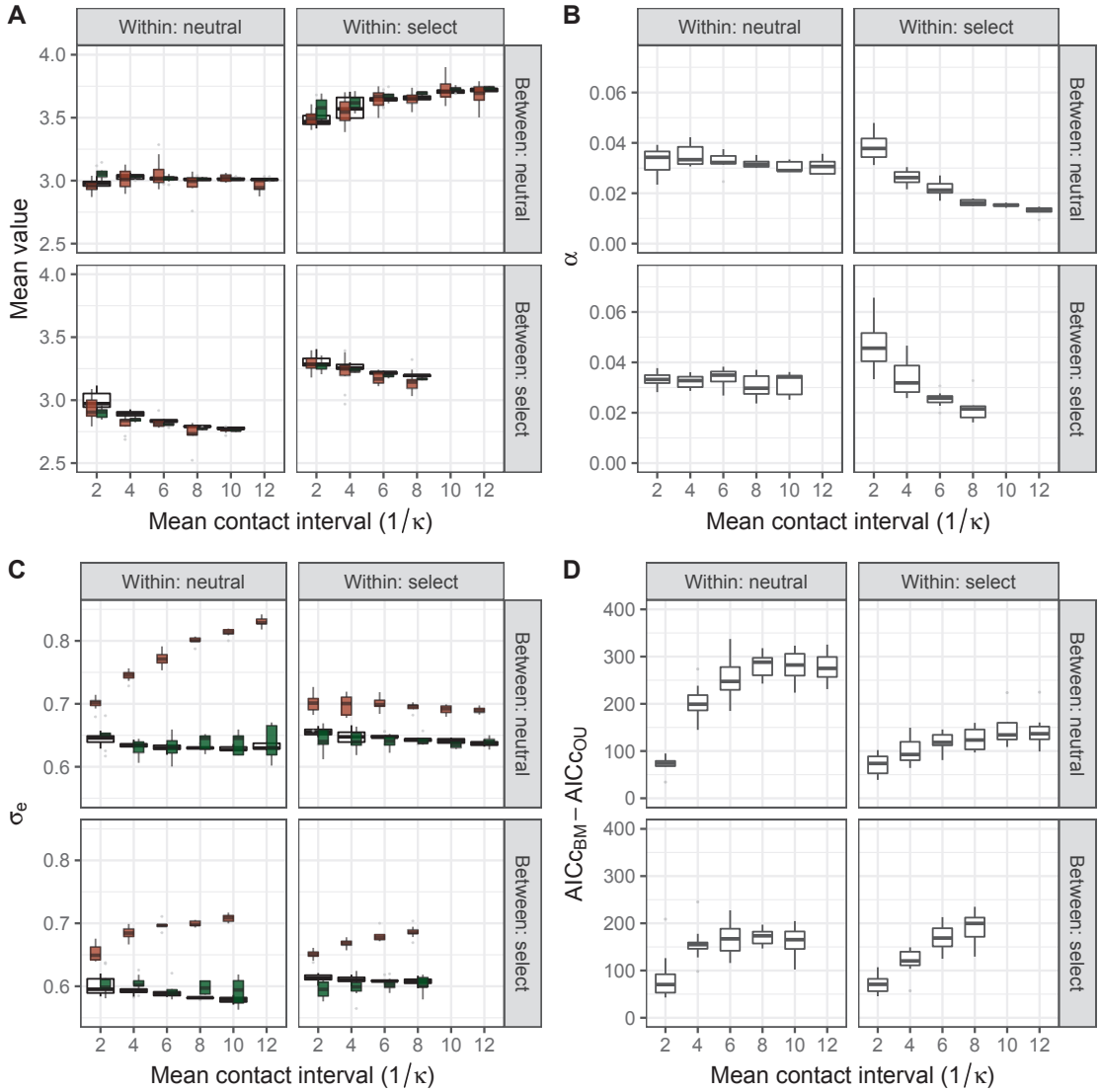


FIG. S6. Details of the PMM and POUMM ML fits to the toy-model simulations. A: comparison beween the true population mean (wide boxes in the background) to the mean-value expected under the PMM method (brown) and the long-term mean value, θ expected under the POUMM method (green); B: Estimates for the parameter α in the toy-model simulations; C: estimates for the parameter σ_e of the PMM (brown) and the POUMM (green) compared to the non-heritable standard deviation estimated from grouping by identical genotype; D: comparison of the corrected Akaike information criterion for the PMM and the POUMM fits - positive values indicate lower (better) AICc for the POUMM method.

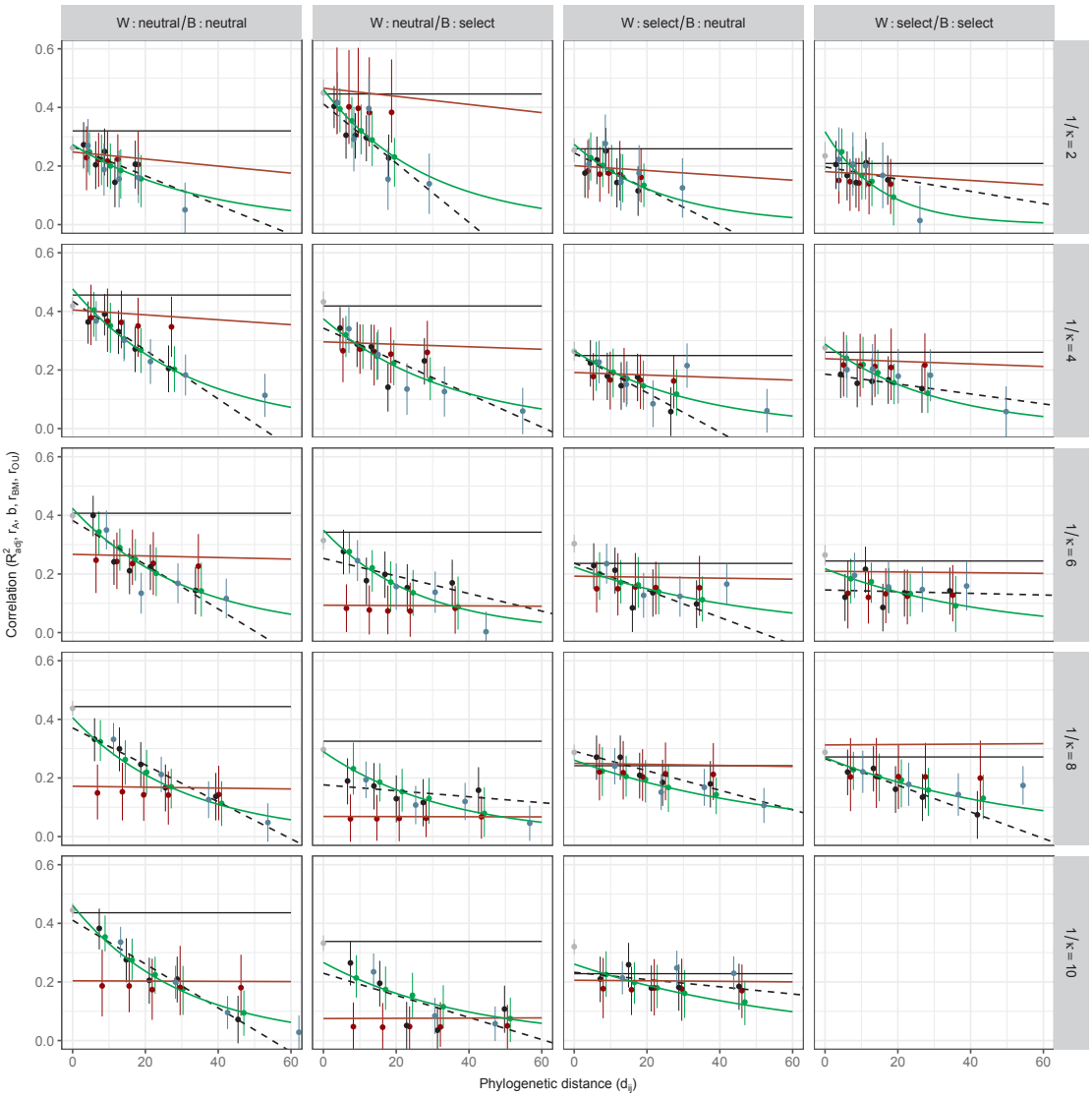


FIG. S7. Correlation in phylogenetic pairs and donor-recipient couples in toy-model simulations. Each panel displays the correlation as a function of d_{ij} in a randomly chosen epidemic for a given scenario and mean contact interval, $1/\kappa$. In each simulated epidemic, we consider the population of the first 10,000 diagnozed individuals. In this population, the exact transmission tree and transmission couples are known. A black horizontal line represents the true value of H^2 measured by the direct estimator R^2_{adj}. Dots and vertical bars display point-estimates and 95% CIs of r_A in the PPs and of b in the donor-recipient couples upon a stratification into quintiles of d_{ij}. Black: r_A in PPs; brown: r_A in PPs after replacing the trait-values simulated under the toy-model with values simulated under the ML fit of the PMM; green: r_A in PPs after replacing the trait-values simulated under the toy-model with values simulated under the ML fit of the POUMM; cadet-blue: b in donor-recipient's; grey (only for $d_{ij}=0$): b_0 in donor-recipient's based on trait-values at moment of infection. A brown and a green line indicate the correlation between tip-pairs in the tree expected under the ML fit of the PMM and POUMM respectively.

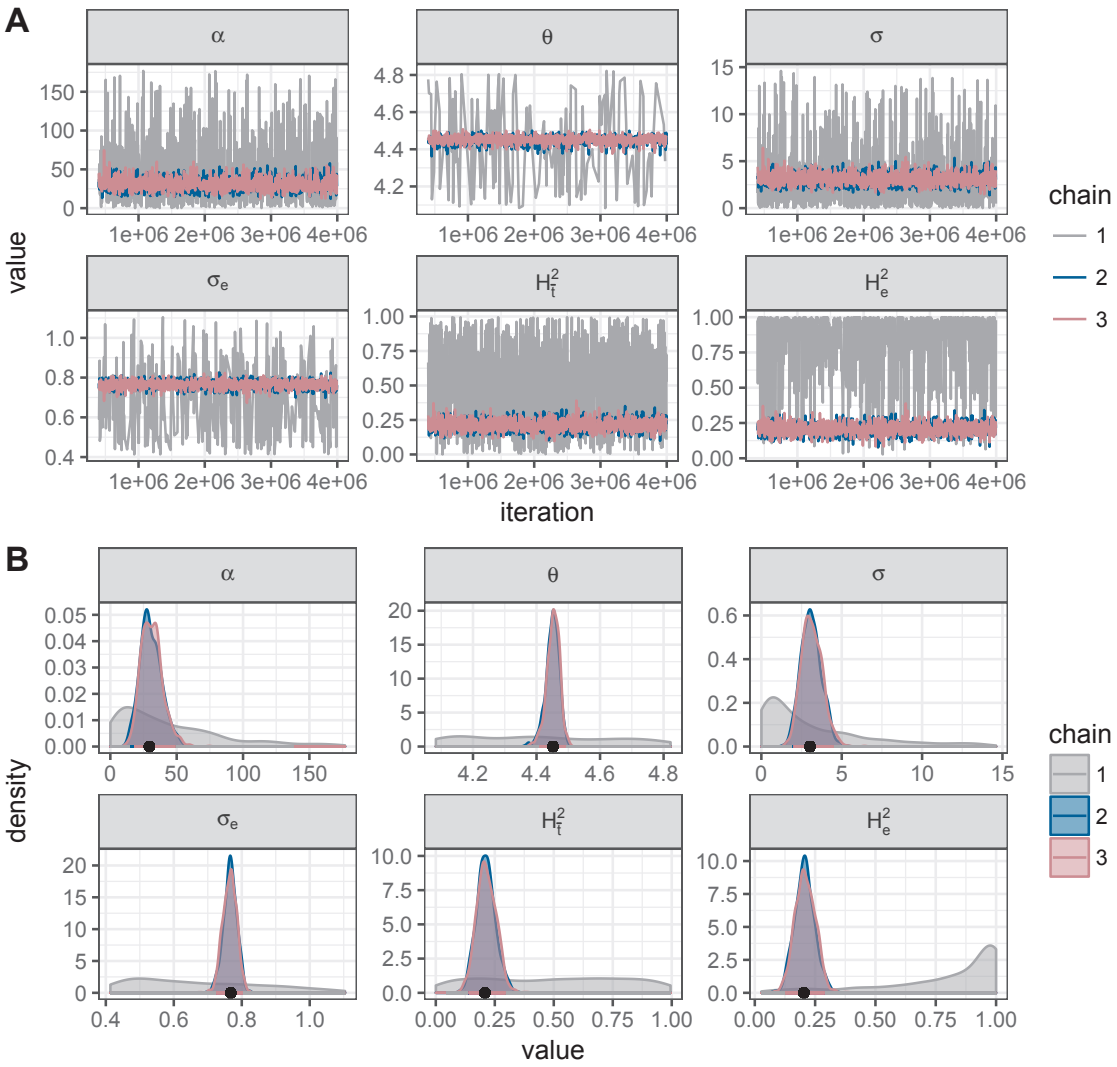


FIG. S8. Trace-plots and posterior densities from the POUMM MCMC-fits to HIV from the UK cohort (8483 patients). Three MCMC chains have been executed: 1 - sampling from the prior distribution; 2 and 3 - sampling from the posterior distribution. (A) Trace-plots - the randomness and the lack of time-correlation in the traces show the correct mixing of the MCMC chain; (B) Inferred posterior densities. The clear distinction between prior and posterior densities proves the presence of informative signal in the data. The match between the densities from chain 2 and 3 proves the convergence of the MCMCs towards the posterior distribution. This convergence was also validated through the Gelman-Rubin statistic being nearly equal to 1 (results not shown).

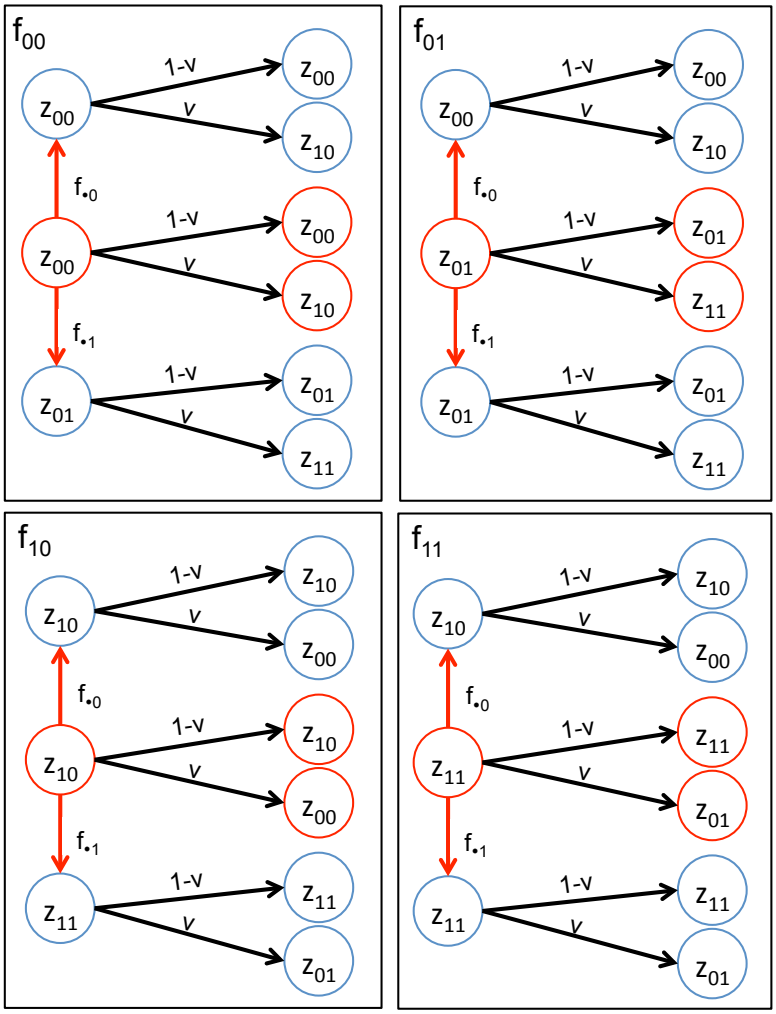


FIG. S9. Expected couples of donor-recipient trait values at $d_{ij}/2$ past transmission Red circles denote donors, blue circles denote recipients. Red vertical arrows denote transmission. Black left-to-right arrows denote mutation during the time from transmission to measurement in the donors and the recipients. The weights above the arrows denote the probability of the transmission or mutation happening. The diagram can be read in the following way (example): at the moment of a generation, a type 00 infected host transmits its pathogen to a susceptible individual of host-type 0 with probability $f_{\cdot 0}$. After the transmission event the strain in each of the two hosts has a chance ν to be substituted by a mutant strain. Thus, the probability of having a donor recipient couple, in which both hosts have a state 00 at the moment of measurement, given that the donor was type 00 at the moment of transmission, is equal to the product $\nu f_{\cdot 0}\nu$. It remains to multiply this by the frequency of encountering a type 00 donor, to obtain the overall probability of the event.

References

Yang, Z. 2006. *Computational Molecular Evolution*. OUP Oxford.

4

DISSECTING HIV VIRULENCE

Published as

Frederic Bertels, Alex Maryel, Gabriel Leventhal, **Venelin Mitov**, Jacques Fellay, Huldrych Günthard, Jürg Boni, Sabine Yerli, Thomas Klimkait, Vincent Aubert, Manuel Battegay, Andri Rauch, Mattias Cavassini, Alexandra Calmy, Enos Bernasconi, Patrick Schmid, Alexandra Scherrer, Viktor Müller, Sebastian Bonhoeffer, Roger Kouyos and Roland Regoës (2017).

Dissecting HIV Virulence: Heritability of Setpoint Viral Load, CD41 T-Cell Decline, and Per-Parasite Pathogenicity. *Molecular Biology and Evolution* 35(1).

This work studies the dynamics of three correlated quantitative traits, characterizing the virulence of an HIV infection. I have the honour to be a co-author of this study of HIV virulence conducted and published by the group of prof. Roland Regös. My contribution has been the adaptation for the purposes of this study of the R-package POUMM, developed originally for the analysis in Chapter 3. In addition I helped in the quantification of confounding factors of HIV virulence, such as patient's sex and age.

Following is the original publication, which appeared in Molecular Biology and Evolution in late 2017.

Dissecting HIV Virulence: Heritability of Setpoint Viral Load, CD4+ T-Cell Decline, and Per-Parasite Pathogenicity

Frederic Bertels,[1] Alex Marzel,[†,2,3] Gabriel Leventhal,[†,‡,1] Venelin Mitov,[4] Jacques Fellay,[5] Huldrych F. Günthard,[2,3] Jürg Böni,[3] Sabine Yerly,[6] Thomas Klimkait,[7] Vincent Aubert,[8] Manuel Battegay,[9] Andri Rauch,[10] Matthias Cavassini,[11] Alexandra Calmy,[12] Enos Bernasconi,[13] Patrick Schmid,[14] Alexandra U. Scherrer,[2,3] Viktor Müller,[15,16] Sebastian Bonhoeffer,[1] Roger Kouyos,[2,3] Roland R. Regoes,[*,1] and the Swiss HIV Cohort Study[17]

[1]Institute of Integrative Biology, ETH Zurich, Zurich, Switzerland

[2]Division of Infectious Diseases and Hospital Epidemiology, University Hospital Zurich, Zurich, Switzerland

[3]Institute of Medical Virology, University of Zurich, Zurich, Switzerland

[4]Department of Biosystems Science and Engineering, ETH Zurich, Zurich, Switzerland

[5]School of Life Sciences, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

[6]Division of Infectious Diseases, Laboratory of Virology, Geneva University Hospital, Geneva, Switzerland

[7]Molecular Virology, Department of Biomedicine – Petersplatz, University of Basel, Basel, Switzerland

[8]Division of Immunology and Allergy, University Hospital Lausanne, Lausanne, Switzerland

[9]Division of Infectious Diseases and Hospital Epidemiology, University Hospital Basel, Basel, Switzerland

[10]Department of Infectious Diseases, Berne University Hospital and University of Berne, Berne, Switzerland

[11]Division of Infectious Diseases, University Hospital Lausanne, Lausanne, Switzerland

[12]HIV/AIDS Unit, Infectious Disease Service, Geneva University Hospital, Geneva, Switzerland

[13]Division of Infectious Diseases, Regional Hospital Lugano, Lugano, Switzerland

[14]Division of Infectious Diseases, Cantonal Hospital St Gallen, St Gallen, Switzerland

[15]Institute of Biology, Eötvös Loránd University, Budapest, Hungary

[16]Evolutionary Systems Research Group, MTA Centre for Ecological Research, Tihany, Hungary

[17]Membership list can be found in the Acknowledgments section

[†]These authors contributed equally to this work.

[‡]Present address: Department of Civil and Environmental Engineering, Massachusetts Institute of Technology (MIT), Cambridge, MA

[*]**Corresponding author:** E-mail: roland.regoes@env.ethz.ch.

Associate editor: Thomas Leitner

Abstract

Pathogen strains may differ in virulence because they attain different loads in their hosts, or because they induce different disease-causing mechanisms independent of their load. In evolutionary ecology, the latter is referred to as "per-parasite pathogenicity". Using viral load and CD4+ T-cell measures from 2014 HIV-1 subtype B-infected individuals enrolled in the Swiss HIV Cohort Study, we investigated if virulence—measured as the rate of decline of CD4+ T cells—and per-parasite pathogenicity are heritable from donor to recipient. We estimated heritability by donor–recipient regressions applied to 196 previously identified transmission pairs, and by phylogenetic mixed models applied to a phylogenetic tree inferred from HIV *pol* sequences. Regressing the CD4+ T-cell declines and per-parasite pathogenicities of the transmission pairs did not yield heritability estimates significantly different from zero. With the phylogenetic mixed model, however, our best estimate for the heritability of the CD4+ T-cell decline is 17% (5–30%), and that of the per-parasite pathogenicity is 17% (4–29%). Further, we confirm that the set-point viral load is heritable, and estimate a heritability of 29% (12–46%). Interestingly, the pattern of evolution of all these traits differs significantly from neutrality, and is most consistent with stabilizing selection for the set-point viral load, and with directional selection for the CD4+ T-cell decline and the per-parasite pathogenicity. Our analysis shows that the viral genotype affects virulence mainly by modulating the per-parasite pathogenicity, while the indirect effect via the set-point viral load is minor.

Key words: **evolution of virulence, disease tolerance, per-parasite pathogenicity, heritability, HIV.**

Article

Introduction

The virulence of an infection is determined by both, the host and the pathogen. One of the most common modulators of virulence is the pathogen load. Higher load often leads to more morbidity, or to faster disease progression or death. Similar to virulence, the load that a pathogen strain attains is also determined by both, the host and the pathogen. In evolutionary ecology, hosts that limit virulence by reducing pathogen load are called "resistant", and pathogen strains that attain a high load in their hosts are often termed "virulent".

But virulence is not completely determined by the pathogen's load alone. There are pathogen-load-independent components, which are again influenced by the host and the pathogen. A host that suffers less than average from being infected by a pathogen and carrying a specific load is called "tolerant" (Caldwell et al. 1958; Schafer 1971; Råberg et al. 2007, 2009; Boots 2008; Boots et al. 2009; Read et al. 2008; Schneider and Ayres 2008; Little et al. 2010; Ayres and Schneider 2012; Medzhitov et al. 2012; Råberg 2014). A pathogen strain that causes less than average virulence attaining a specific load is said to have a low "per-parasite pathogenicity" (Råberg and Stjernman 2012; Råberg 2014). Figure 1A displays these virulence components diagrammatically.

How can pathogen-load-independent components of virulence be determined? To identify these components, "excess virulence" needs to be measured, that is by how much virulence differs from what is expected for a specific pathogen load. Statistically speaking, "excess virulence" is the residual virulence after adjusting for differences in pathogen load. This adjustment can be visualized on fitness-versus-pathogen-load plots (fig. 1B). On such a plot, host types with differing levels of disease tolerance are characterized by different tolerance curves that depict the relationship between virulence and pathogen load (see fig. 1B bottom left). The steepness of this curve is inversely related to disease tolerance.

Once tolerance curves for different host types are determined, the per-parasite pathogenicity will manifest itself as a yet unexplained deviation from the tolerance curve. In other words, varying degrees of per-parasite pathogenicity will lead to residual excess virulence that is not explained by host factors. Figure 1C shows how two pathogen strains differing in their per-parasite pathogenicity will scatter around the tolerance curves of two host types.

HIV infection provides an ideal example to illustrate this decomposition of virulence. In this infection, CD4+ T cells—the target cells of the virus—continuously decline from a level of ~1,000 cells per microliter blood. A CD4+ T-cell count below 200 cells per microliter blood is one of the defining characteristics of AIDS. The decline rate of the CD4+ T cells is a well-established surrogate for the rate of disease progression (Phillips et al. 1991), that is virulence. It has the advantage that it can be determined from clinical samples spanning less than one year of monitoring an HIV infected individual, whereas the direct observation of disease progression requires many years. The existence of such a well-established, quantitative measure of virulence makes HIV infection unique.

The decomposition of virulence relies on its relation with pathogen load. During HIV infection, the viral load rises and peaks a few weeks after infection, and subsequently settles at a fairly stable level that is maintained for many years—the so-called set-point viral load. This set-point viral load represents a good measure of pathogen load required for the decomposition of virulence, and it is associated with the rate of progression toward disease and death (Mellors et al. 1996). However, the correlation between the set-point viral load and the decline of the CD4+ T cells—a good proxy of the rate of disease progression—is not very strong: R^2 values were found to be between 0.05 and 0.08 in American cohorts (Rodriguez et al. 2006), and 0.05 for the population studied here (Regoes et al. 2014) (although the correlation between set-point viral load and survival time has been reported to be higher, Arnaout et al. 1999). While this weak correlation may be in part the result of measurement error it suggests that there are factors influencing virulence other than the set-point viral load.

In the context of HIV infection, examples for variation in all of the four virulence components exist. First, the set-point viral load differs by three orders of magnitude between HIV infected individuals, ranging from 10^3 to 10^6 RNA copies per ml plasma (Mellors et al. 1996; Raboud et al. 1996; Deeks et al. 1997). This variation is associated with the rate of disease progression (Mellors et al. 1996). Recent heritability studies (as reviewed in Fraser et al. 2014) have shown that a fraction of the variation in the set-point viral load can be attributed to the viral genotype infecting the host.

Second, human genes conferring "host resistance" in the sense of evolutionary ecology (see above) have been identified: Individuals who carry protective HLA-B alleles have lower set-point viral loads, and progress to disease at a slower rate than people without these alleles (Goulder and Watkins 2008; Fellay et al. 2007). Across primates, species-specific restriction factors can limit the load of Human and Simian Immunodeficiency Viruses (SIVs; Kirchhoff 2009; Zheng et al. 2012).

Third, in the context of immunodeficiency viruses, there are examples for variation in pathogen-load-independent virulence components. SIV infection in natural hosts, such as the sooty mangabeys, is avirulent despite the high set-point viral loads SIV attains in these hosts (Chakrabarti 2004; Chahroudi et al. 2012). In contrast, SIV infection in nonnatural hosts, such as the rhesus macaque, leads to an AIDS-like disease (Chakrabarti 2004; Chahroudi et al. 2012). Thus, natural hosts tolerate the infection without becoming sick. There is also variation in tolerance to HIV infection in humans associated with age and human leukocyte antigens (HLA; Regoes et al. 2014).

Lastly, there is also variation in per-parasite pathogenicity across viral strains. For example, there is evidence that HIV-1 subtype D leads to faster disease progression than subtype A even though both subtypes attain similar set-point viral loads (Baeten et al. 2007). In other words, subtype A and D differ in their per-parasite pathogenicity.

Previously, we investigated if humans display variation in disease tolerance against HIV (Regoes et al. 2014). We found

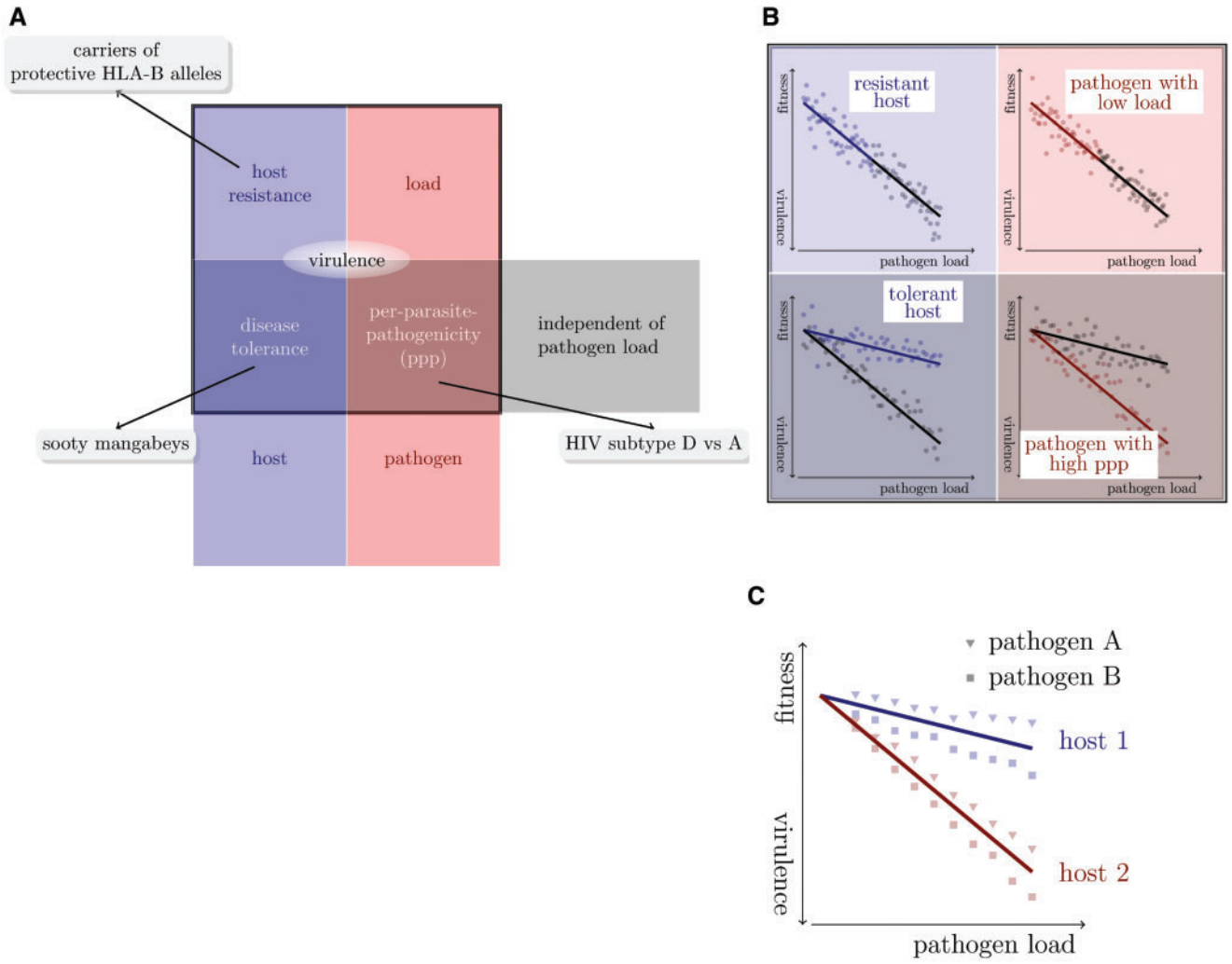


FIG. 1. Dissecting virulence. (*A*) Systematics of virulence components. Each component can be a trait of either the pathogen or the host, and depend or be independent of the load of the pathogen. (*B*) Formally, virulence can be dissected using fitness-versus-pathogen-load plots. (In these plots, host fitness is inversely correlated with virulence.) Adapted from figure 1 in Råberg (2014). (*C*) In multi-host multi-pathogen systems, virulence components can be disentangled by first defining host-type-specific tolerance curves. Pathogens differing in their per-parasite pathogenicity will then fall on different sides of these tolerance curves. In the example shown, pathogen B has a higher per-parasite pathogenicity than pathogen A.

that younger individuals and HLA-B heterozygotes are more tolerant, and could link the variation in tolerance to HLA-B genotype. In this previous study, we did, however, not investigate the potential impact of the virus genotype on virulence.

How can one investigate if the virus genotype influences a host–pathogen trait? One way is to study associations between genetic polymorphisms of the virus and the trait, as is done, for example, in genome-wide association studies (Bartha et al. 2013). An alternative approach is to estimate the heritability of the trait. If a trait is heritable, that is similar between similar viral genotypes, it must, at least in part, be determined by viral genes.

The influence of the virus genotype on the set-point viral load has been the focus of many research groups, including ours, over the past years. Most studies determined the heritability of the set-point viral load (Alizon et al. 2010; Hollingsworth et al. 2010; Müller et al. 2011; Hodcroft et al. 2014; Fraser et al. 2014), while others investigated associations between genetic polymorphisms of the virus and the trait (Bartha et al. 2013). There is a consensus that set-point viral load is heritable, although there is some controversy on the numerical value of the heritability.

In this study, we investigate the influence of the HIV genotype on overall virulence, as measured by the CD4+ T-cell decline, and its pathogen-load-dependent and -independent components, the set-point viral load and per-parasite pathogenicity, respectively. We determine the influence of the viral genotype by estimating the "heritability" of these traits, measuring how similar the trait values are in the donor and in the recipient. To this end, we use data from the Swiss HIV Cohort Study. For our analysis, we selected cohort participants, for whom we could determine the set-point viral load and the decline of CD4+ T lymphocytes—an established proxy for virulence in HIV infection. As a surrogate for the per-parasite pathogenicity we use the residuals from

previously determined tolerance curves (Regoes et al. 2014) as described in Materials and Methods.

Our analysis confirms that set-point viral load is heritable. We further provide evidence the virulence of HIV infection, as measured by the decline of CD4+ T lymphocytes, is heritable. Lastly, we find evidence for the heritability of the per-parasite pathogenicity, the pathogen-load-independent component of virulence. Our results are therefore consistent with the notion that the virus genotype affects virulence in HIV infection both via the viral load, and via viral-load-independent mechanisms.

Results

Heritability of Set-Point Viral Load Confirmed

The heritability of the set-point viral load has previously been estimated from data of the Swiss HIV cohort study (Alizon et al. 2010) and from data of other cohorts (Hollingsworth et al. 2010; Müller et al. 2011; Hodcroft et al. 2014). The methods differed across these studies.

To test the conclusion of these studies, we applied donor–recipient regressions and the phylogenetic mixed models to the set-point viral loads from the 2014 individuals we included in the present study. The donor–recipient regressions were applied to 196 previously determined transmission pairs (Kouyos et al. 2014), whereas the phylogenetic methods were applied to a phylogenetic tree that was constructed from *pol* gene sequences (see Materials and Methods). Since set-point viral loads were significantly associated with sex, age at infection, and were higher in men who have sex with men, we also estimated the heritability of adjusted set-point viral loads as defined in Materials and Methods.

Across all the methods we use, the estimates for heritability range from 8% to 29% (see table 1, fig. 2*B* and supplementary fig. S2*B*, Supplementary Material online). These estimates are all significantly larger than zero, except for the adjusted set-point viral load using the phylogenetic mixed-model that assumes neutral evolution. These results add to the growing consensus that set-point viral load is heritable.

Interestingly, assuming stabilizing selection on the set-point viral load in our phylogenetic analysis led to a significantly better fit to the data than assuming neutral drift (Likelihood ratio test: $P = 1.2 \times 10^{-4}$ for unadjusted and $P = 8.8 \times 10^{-6}$ for adjusted set-point viral loads). Thus, the estimates for the heritability of the set-point viral load with the best statistical support are 26% and 29% without and with adjustment for cofactors, respectively. Both of these estimates are significantly different from zero.

The optimal trait value θ of the Ornstein–Uhlenbeck process is estimated to be $10^{4.0}$ RNA copies per milliliter plasma for the unadjusted set-point viral load (95% CI: $10^{1.6}$–$10^{4.3}$ RNA copies per milliliter plasma), very close to the mean of the set-point viral load in our study population of $10^{4.2}$ RNA copies per milliliter plasma (see table 2). The parameter measuring the strength of selection, α is high: 32.7 (95% CI: 0.03–57.6) and 39.4 (95% CI: 6.1–68.1) for unadjusted and adjusted set-point viral load, respectively (see table 2). These parameter estimates are consistent with strong stabilizing selection around the current mean trait value.

Evidence for the Heritability of CD4+ T-Cell Decline

The set-point viral load is an important determinant of CD4+ T-cell decline, and it is heritable. We therefore expect the CD4+ T-cell decline to be also heritable "by association".

Table 1. Heritability Estimates for Set-Point Viral Load (spVL), CD4+ T-Cell Decline (ΔCD4), and Per-Parasite Pathogenicity (ppp) Based on the Phylogenetic Mixed Models Assuming Brownian Motion-type Trait Evolution (PMM) or Trait Evolution According to the Ornstein–Uhlenbeck Process (POUMM).

	PMM	POUMM
ΔCD4 (unadjusted)	**25% (9%–40%)**	**17% (6%–29%)**
ΔCD4 (adjusted)	**24% (7%–39%)**	**17% (5%–30%)**
spVL (unadjusted)	**12% (2%–28%)**	**26% (8%–43%)**
spVL (adjusted)	**8% (0%–26%)**	**29% (12%–46%)**
ppp	**22% (5%–39%)**	**17% (4%–29%)**

NOTE.—95% confidence intervals are given in brackets.

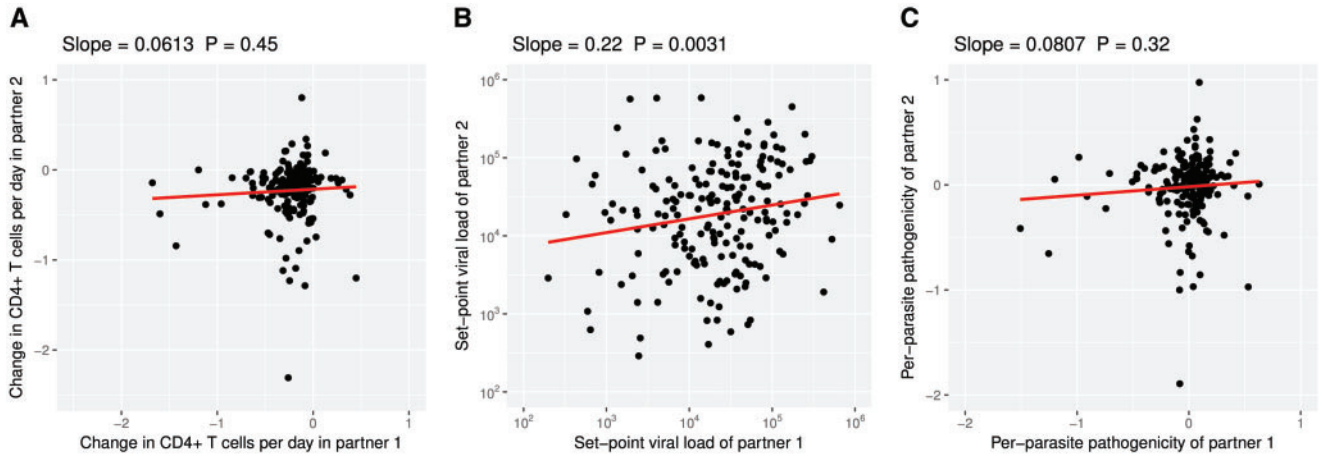


FIG. 2. Heritability estimates from donor–recipient regressions. For these regressions we plotted the trait values for each partner ("partner 1" and "partner 2") in the transmission pairs onto the *x*- and *y*-axes. Since we do not know the direction of the transmission in the pairs, the assignment of the partners to either *x*- or *y*-axis is random. (See also supplementary fig. S2, Supplementary Material online.)

Table 2. Estimates of the POUMM Parameters Related to Selection for Set-Point Viral Load (spVL), CD4+ T-Cell Decline (ΔCD4), and Per-Parasite Pathogenicity (ppp).

	α	θ	Population Mean
ΔCD4 (unadjusted)	4.1 (0.5–10.8)	−1.15 (−2.43, −0.29)	−0.20
ΔCD4 (adjusted)	3.8 (0.4–10.5)	−0.93 (−2.30, −0.06)	−0.04
spVL (unadjusted)	32.7 (0.03–57.6)	4.0 (1.6, 4.3)	4.2
spVL (adjusted)	39.4 (6.1–68.1)	−0.03 (−0.17, 0.10)	−0.04
ppp	3.9 (0.5–10.4)	−0.89 (−2.17, −0.09)	0.00

NOTE.—95% confidence intervals are given in brackets.

So far, however, there has been no evidence for the heritability of the CD4+ T-cell decline or HIV virulence.

To assess the heritability of the CD4+ T-cell decline, we applied the same methods as for the set-point viral load. Because the CD4+ T-cell decline was associated significantly only with the age at infection we also conducted the analyses with age-adjusted CD4+ T-cell declines as defined in Materials and Methods.

The donor–recipient regression (fig. 2A) results in a heritability estimate, which is not significantly different from zero for both unadjusted and age-adjusted CD4+ T-cell declines. This is likely the result of the low number of individuals in the transmission pairs (2×196), which limits the statistical power of the donor–recipient regressions.

Using the phylogenetic mixed models, however, we can incorporate all 2014 individuals of our study population, and obtain heritability estimates significantly larger than zero. Assuming neutral trait evolution (PMM) yields 25% and 24% for unadjusted and adjusted CD4+ T-cell declines, respectively. With trait evolution according to the Ornstein–Uhlenbeck process we get 17% irrespective of any adjustment (see table 1).

Again, assuming trait evolution according to an Ornstein–Uhlenbeck process has more statistical support than Brownian motion trait evolution (Likelihood ratio test: $P = 6.9 \times 10^{-8}$ for unadjusted and $P = 2.6 \times 10^{-5}$ for adjusted CD4+ T-cell declines). But, unlike for the set-point viral load, the estimate of the optimal trait value θ (-1.15 per day, 95% CI: -2.43 to -0.29 per day) is significantly below the mean of the CD4+ T-cell declines in our study population (-0.20 per day), and the estimated strength of selection α is an order of magnitude lower than that for the set-point viral load (4.1 and 3.8 for unadjusted and adjusted CD4+ T-cell decline, respectively). See table 2 for all selection related parameter estimates for this trait. These parameters characterize directional, rather than stabilizing selection, consistent with a slow time trend towards higher virulence.

Evidence for the Heritability of the Per-Parasite Pathogenicity

Lastly, we investigated if there is any evidence for the heritability of the per-parasite pathogenicity. The per-parasite pathogenicity is the component of virulence, which is determined by the pathogen genotype and independent of the pathogen load. Formally, we determine per-parasite pathogenicity by calculating the residual of the regression of the CD4+ T-cell decline against the set-point viral load adjusted for the age of the infected individual (see Materials and Methods and fig. 3).

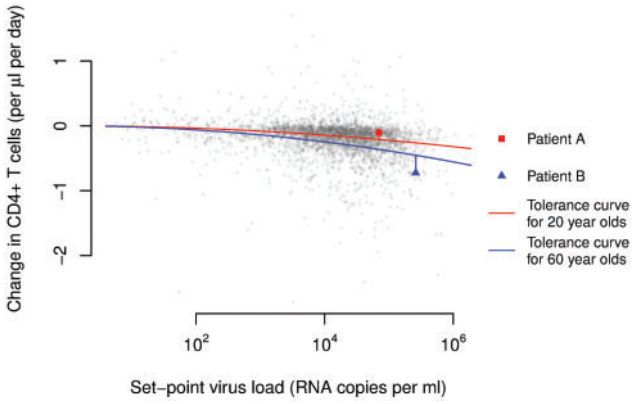


FIG. 3. As a surrogate for the per-parasite pathogenicity we use the residuals from age-adjusted tolerance curves. In the graph, we plotted each individual's CD4+ T-cell decline versus his/her set-point viral load. The red and blue curves show the average relationships between these two measures in the groups that were 20 and 60-years-old at the time of their infection, respectively. These age-adjusted tolerance curves were determined previously (Regoes et al. 2014). The red square and blue triangle highlight two individuals with an age at infection of 20 and 60 years, respectively.

A donor–recipient regression (fig. 2C) yields an estimate, which is not significantly different from zero, again likely due to the low number of transmission pairs. Using phylogenetic mixed models, we estimate a statistically significant heritability of 22% assuming Brownian motion-type trait evolution, and 17% for trait evolution according to the Ornstein–Uhlenbeck process (see table 1).

As for the set-point viral load and the CD4+ T-cell decline, assuming trait evolution governed by an Ornstein–Uhlenbeck process led to a significantly better fit than evolution according to Brownian motion (Likelihood ratio test: $P = 2.2 \times 10^{-6}$). The pattern of selection most consistent with the evolution of per-parasite pathogenicity is directional selection as for the CD4+ T-cell decline: the optimal trait value is estimated to be significantly lower than the population mean, and the strength of selection is weak (see table 2). This suggests that there is a slow time trend of increasing per-parasite pathogenicity.

Testing for Genetic Associations with Per-Parasite Pathogenicity

It is known that HIV-1 subtype D has a higher per-parasite pathogenicity than subtype A. These intersubtype difference in disease progression have been mapped to the *pol* gene and were associated with replicative capacity (Ng et al. 2014). In particular, a valine instead of an isoleucine at position 62 and 64 in the protease (I62V and I64V), and a proline instead of an

alanine at position 272 in the reverse transcriptase were found to be associated with high replicative capacity (Ng et al. 2014).

We tested if these amino acid polymorphisms are associated with set-point viral load, CD4+ T-cell decline, and per-parasite pathogenicity within HIV-1 subtype B. Indeed isoleucine and valine at positions 62 and 64 of the protease, and alanine and proline at position 272 in the reverse transcriptase were the most prevalent amino acids in our study population. The set-point viral load was not associated with any of these polymorphisms. Before adjustment for multiple comparisons, the association of proline at position 272 in the reverse transcriptase with CD4+ T-cell decline and per-parasite pathogenicity reach a significance level of $P = 0.014$ and $P = 0.020$, respectively, which is not significant after adjusting the P-values for three comparisons.

Discussion

In this study, we confirmed the heritability of the set-point viral load. Further, we report one of the first pieces of evidence for the heritability of the CD4+ T-cell decline, a surrogate of virulence. We also found support for the hypothesis that the pathogen-load-independent virulence component is heritable. Lastly, the evolution of these three traits is significantly better described by the Ornstein–Uhlenbeck process than by Brownian motion.

Our study confirms previous studies that established the heritability of the set-point viral load in HIV infection (as reviewed by Müller et al. 2011; Fraser et al. 2014). In particular, our estimates are consistent with those from a donor–recipient regression in Hollingsworth et al. (2010), and two recent studies applying phylogenetic mixed models based on the Ornstein–Uhlenbeck process by Mitov and Stadler (2016) and Blanquart et al. (2017). Our analysis is also consistent with the study by Hodcroft et al. (2014) that reported a low heritability of 5.7% of the set-point viral load adjusted for covariates and assuming Brownian trait evolution. If we adjust for covariates and assume Brownian trait evolution, we obtain a heritability estimate of 8% that is not significantly different from zero. Assuming trait evolution according to the Ornstein–Uhlenbeck process, however, provides a significantly better fit to the adjusted set-point viral load data and yields a heritability estimate of 29%.

We find clear evidence for the heritability of the CD4+ T-cell decline. As the level of CD4+ T cells is a defining characteristic of clinical AIDS, the CD4+ T-cell decline is a good surrogate of virulence of HIV infection. Although the potential heritability of the rate of decline of CD4+ T cells has been investigated previously (Alizon et al. 2010), it was found to be not significantly different from zero. We attribute this discrepancy to the low sample size of the earlier study. In contrast to the 2014 individuals in our study population, Alizon et al. (2010) had enrolled only 1,100 and investigated the heritability only in subpopulations consisting of a few hundred individuals.

The recent study by Blanquart et al. (2017) also reports heritability of the CD4+ T-cell decline. For the HIV-1 subtype B, they estimate a heritability of 11% ranging from 0% to 19%. This estimate is within the 95% confidence intervals of our estimate, and our estimate of 17% is within the 95% confidence intervals of their estimate. Unlike our analysis, Blanquart et al. (2017) did not find support for the Ornstein–Uhlenbeck over the Brownian motion trait evolution model. Their estimate of the strength of selection on the CD4+ T-cell decline is 0.095 whereas we estimate 3.8. This may be due to the lower sample size of 1,170 individuals in the study by Blanquart et al. (2017). It may also be the result of differences in the inclusion criteria: Blanquart et al. (2017) included individuals with five CD4+ T-cell measures between the time of their first positive HIV test and the beginning of treatment, while we require only three CD4+ T cells but in a more stringent time window that excludes the first 90 days after the estimated time of infection and time points, after the CD4+ T-cell count fell below 100 cells per microliter blood. Differences in the tree reconstruction algorithm, however, are unlikely to explain the discrepancy. If we reconstruct the phylogenetic tree with RaxML (Stamatakis 2006b), which is closely related to ExaML (Kozlov et al. 2015) that Blanquart et al. (2017) used, we obtain very similar parameter estimates (see supplementary table S1, Supplementary Material online). The heritability estimates based on the FastTree and RaxML trees are most discrepant for the adjusted set-point viral load when we use POUMM (29% vs. 34% or 38%), but, because of the large confidence intervals, these estimates are not statistically different. With the RaxML trees, we also find that Ornstein–Uhlenbeck trait evolution has higher statistical support.

We also provide evidence for the heritability of the per-parasite pathogenicity. This trait describes the pathogenic potential of a viral strain that is independent of the load the strain attains in its host. We approximated this trait as the deviation of the CD4+ T-cell decline observed in an individual from that predicted on the basis of the observed set-point viral load and the age of the infected host (see Materials and Methods). In addition to being determined by per-parasite pathogenicity, this deviation could be affected by further host factors, other sources of biological variation, and, of course, measurement noise, and should therefore be considered only as a surrogate for the per-parasite pathogenicity of a strain. It is important to note, however, that the uncertainties surrounding the quantification of the per-parasite pathogenicity make it more difficult to establish the heritability of this trait. The fact that we found evidence for heritability means that the signal in our surrogate measure of the per-parasite pathogenicity is not completely clouded by factors, for which we could not account.

For all three traits we considered, the Ornstein–Uhlenbeck trait evolution model has the best statistical support. For the set-point viral load the strength of selection is estimated to be high, and the inferred optimal value of the trait is close to the mean of the set-point viral load in our study population. Thus, set-point viral load is under significant stabilizing selection. The CD4+ T-cell decline and the per-parasite pathogenicity, on the other hand, are not under strong stabilizing selection but directional selection—a scenario that the

Ornstein–Uhlenbeck trait evolution model can capture with low selection strength α and an optimal trait value θ significantly different from the population mean. The parameter estimates of the Ornstein–Uhlenbeck model for these two traits are consistent with a slow but significant increase in HIV virulence over the past two decades (Pantazis et al. 2014). It is well-recognized that the Ornstein–Uhlenbeck trait evolution model can accommodate these two distinct patterns of selection. The difference in the nature of selection between set-point viral load and the other two traits is also the reason behind the bias in the heritability estimates based on Brownian motion trait evolution: the heritability of the set-point viral load is underestimated, whereas those of the other two traits are slightly overestimated. This conclusion is in agreement with simulation studies of this bias (Mitov and Stadler 2017).

Intuitively, the heritability of the CD4+ T-cell decline should be the combination of set-point viral load dependent and independent components of this trait. The heritability estimates we obtained conform surprisingly well with this expectation. The estimate of the heritability of the CD4+ T-cell decline with the highest statistical support is 17%. The adjusted set-point viral load has a heritability of 29%. To approximate to what extent the heritability of the set-point viral load will trickle through to that of the CD4+ T-cell decline, we need to factor in the correlation between these two traits. In our study population, the fraction of the variation in the CD4+ T-cell decline explained by the set-point-viral load is $R^2 = 0.057$, in very close agreement with estimates of this quantity in other study populations (Rodriguez et al. 2006). Thus, $\sim 29\% \times 0.057 = 1.6\%$ of the 17.4% of the heritability in the CD4+ T-cell decline are due to the set-point viral load. The remainder—$\sim$16%—should be independent of the set-point viral load. This agrees well with our estimate of the heritability of the per-parasite pathogenicity of 17%, especially given the uncertainty in all of these estimates.

The heritability of the per-parasite pathogenicity means that there are viral genes that influence the CD4+ T-cell decline in ways that do not depend on the viral load. Generally, one conceivable such mechanism could be that viral genotypes with high per-parasite pathogenicity elicit ineffective immune responses that, rather than reducing viral load, accelerate CD4+ T-cell decline. In terms of specific viral factors influencing per-parasite pathogenicity, studies that compare the disease course across HIV-1 subtypes are illuminating. In particular, the coreceptor usage (Daar et al. 2007) and *pol* replicative capacity (Barbour et al. 2004; Goetz et al. 2010; Ng et al. 2014) have been found to be associated with the rate of disease progression independently of the set-point viral load. Ng et al. (2014) identified three amino acid polymorphisms that are associated with replicative capacity. We tested if these polymorphisms are associated with per-parasite pathogenicity, CD4+ T-cell decline and set-point viral load. While we could not establish any clear-cut association after adjusting for multiple comparisons, a proline at position 272 of the reverse transcriptase was the most promising candidate. The sample size for this test was lower ($n = 1222$) than the size of our study population ($n = 2014$) because the sequence region containing position 272 in the reverse transcriptase was not available in all individuals. This polymorphisms should be tested in the future for an association with CD4+ T-cell decline and per-parasite pathogenicity in a larger study population.

Previously, Pagel's λ was also employed to estimate the heritability of the set-point viral load from a phylogenetic tree of HIV *pol* sequences (Alizon et al. 2010). Leventhal and Bonhoeffer (2016), however, have argued recently that Pagel's λ implicitly assumes the trees to be ultrametric. Thus, for nonultrametric phylogenetic trees, such as the one we analyzed, this essential assumption of Pagel's λ is violated, and this method should therefore not be used. We nevertheless provide heritability estimates for comparison in supplementary figure S1, Supplementary Material online. With Pagel's λ the set-point viral load and the CD4+ T-cell decline are also found to be significantly heritable.

The heritability of traits from donor to recipient is sometimes interpreted as the quantitative measure of the extent to which the trait is "under the control of the virus". It has been pointed out that the viral genome carries an imprint of past environments, in particular the immune responses experienced in former hosts (Bartha et al. 2013, 2017; van Dorp et al. 2014; Carlson et al. 2014 2016). Thus, separating virus from host effects is challenging. What is undisputed is that the trait in the new host is in part encoded in the viral genome. Information on associations with viral genes is very valuable, especially for traits that cannot be determined instantaneously, such as the CD4+ T-cell decline.

In summary, we presented a comprehensive evolutionary analysis of the components of HIV virulence. We established that viral load dependent and independent virulence components, as well as overall virulence are heritable. This strongly suggests that these virulence components are, at least in part, encoded in the viral genome. Future research will need to identify the specific genetic polymorphisms associated with these virulence components.

Materials and Methods

Study Population

We analyzed a subset of the individuals from the Swiss HIV Cohort Study (www.shcs.ch; last accessed September 24, 2017; Schoeni-Affolter et al. 2010). This study has enrolled >19,000 HIV-infected individuals to date, which constitutes >72% of all patients receiving antiretroviral therapy in Switzerland, and is therefore highly representative. The viral load and CD4+ T-cell count of each enrolled individual are determined approximately every three months. In some of these individuals, the *pol* gene of the virus was sequenced. The *pol* gene encodes viral enzymes important for viral replication within its target cell, most notably the reverse transcriptase, the integrase, and the protease, and contains many of the clinically relevant resistance mutations.

The study population of the present study consists of a subset of the study population analyzed in a previous study (Regoes et al. 2014). In this previous study, we had included 3,036 HIV-1 infected individuals, for whom viral load measurements and CD4+ T-cell counts were available to reliably

estimate the set-point viral load and CD4+ T-cell decline. We restricted our analysis to data obtained before antiretroviral treatment. Furthermore, we excluded the primary and late phases of the infection by discarding measurements during the first 90 days after the estimated date of infection and measurements obtained when the CD4+ T-cell count was below 100 per microliter blood. Lastly, individuals were included if they had at least two viral load measurements and three CD4+ T-cell measurements that were at least 180 days apart.

For the present study, we selected 2014 individuals of the 3,036 individuals enrolled previously. Individuals were included if the *pol* gene of their virus had been sequenced. The genetic information of the virus was necessary for the present study to infer the evolutionary history and investigate patterns of heritability.

Pol sequence information was obtained from the SHCS genotypic drug resistance database. Sequences are stored in a central database (SmartGene; Integrated Database Network System version 3.6.13). All laboratories perform population-based sequencing (von Wyl et al. 2007; Yang et al. 2015). The drug resistance database includes, in addition to the routinely collected samples, over 11,000 samples from the biobank analyzed by systematic retrospective sequencing (Yang et al. 2015; Schoeni-Affolter et al. 2010). The individuals in our study population belong to the following risk groups: men having sex with men—972 (48%), heterosexuals—435 (21.5%), intravenous drug users—365 (18%), and others—252 (12.5%).

The SHCS, enrolling HIV-infected adults aged over 16-years-old, has been approved by ethics committees of all participating institutions. The data collection was anonymous and written informed consent was obtained from all participants (Schoeni-Affolter et al. 2010).

Set-Point Viral Loads and CD4+ T-Cell Declines

For each individual enrolled in our study, the set-point viral load had been determined in a previous study (Regoes et al. 2014) as the mean of the logarithm to the base 10 of the eligible viral load measurements in each individual. Nondetectable viral loads had been set to half the detection limit. The rate at which the CD4+ T cells change per day had previously been estimated as the slope in a linear regression of CD4+ T-cell counts in an individual against the date, at which they were determined. The rate of change in CD4+ T cells is inversely related to virulence.

Set-point viral loads and CD4+ T-cell declines were adjusted for potential covariates by regressing them against sex, age at infection, risk group and ethnicity. Once significant covariates were identified, adjusted traits were defined as the residuals of a regression with these covariates. Subsequent analyses were then conducted with the residuals.

The inclusion criteria, calculation of set-point viral load and CD4+ T-cell decline, as well as the model fitting and comparisons had been implemented and performed in the R language of statistical computing (R Core Team 2013).

Per-Parasite Pathogenicity

Per-parasite pathogenicity is defined as the pathogenic potential of a pathogen strain adjusted for its load and host factors that are associated with pathogen-load independent virulence components.

To derive a proxy for the per-parasite pathogenicity of a strain, we first determined the relationship between pathogen load and pathogenicity—called the tolerance curve—for a given host type. In our previous study (Regoes et al. 2014), we found that the age, at which the host was infected, was associated very strongly with the slope of the tolerance curve. Therefore, we determined the tolerance curve specific for the age of the host that harbors the pathogen strain. We did not account for host factors other than age, such as, for example, HLA-B genotype or homozygosity, because this information of host genotype is lacking for the majority of our study population.

In a next step, we predicted the CD4+ T-cell decline we should observe given the set-point viral load that this strain attains in the host. Lastly, we calculated by how much the observed CD4+ T-cell decline deviates from this prediction (see fig. 3).

In our conceptualization, tolerance is measured by the parameter characterizing the relationship between the CD4+ T-cell decline and the set-point viral load. Different levels of per-parasite pathogenicity, on the other hand, manifest themselves as a deviation of an individual's CD4+ T-cell decline and set-point viral load from the tolerance curve. Formally, this procedure amounts to regressing CD4+ T-cell decline against the set-point viral load, adjusting for the age at infection, and calculating the residual of an individual's trait from the regression line. We denote this proxy as ppp. Lower and more negative values of this quantity are associated with faster disease progression.

Transmission Pairs and Phylogenetic Tree

The transmission pairs were identified as monophyletic clusters on a previous HIV transmission tree (Kouyos et al. 2014). Of these previously established transmission pairs, 196 were present in our study population. The direction of transmission cannot be inferred in these pairs.

The reconstruction of the phylogenetic tree relies on *pol* gene sequencing of the virus carried by the study subjects. In particular, we had sequences of *pol* extending over the HXB2 positions 2253–3870, comprising the protease and the reverse transcriptase. All sequences were initially aligned to an HXB2 reference genome (http://www.ncbi.nlm.nih.gov/nuccore/K03455.1; last accessed September 24, 2017) using MUSCLE (Edgar 2004). We selected the earliest sequence if more than one sequence was available for a person.

To reconstruct the evolutionary history, we first removed insertions relatively to HXB2. To exclude signatures of parallel evolution due to drug pressure that can distort the inferred evolutionary history, we further removed drug resistance mutations according to the databases of Stanford (http://hivdb.stanford.edu/; last accessed September 24, 2017) and the International Antiviral Society (https://www.iasusa.org/; last accessed September 24, 2017). We used Gblocks to refine the alignment. The final number of positions was 1106.

We constructed the phylogenetic tree using FastTree (Version 2.1.8 SSE3, OpenMP; Price et al. 2010). We used a maximum-likelihood-based inference using a Generalized Time-Reversible evolutionary model and a CAT model (Stamatakis 2006a) with 20 discrete evolutionary rate categories. We use the most rigorous and time-consuming FastTree parameters (FastTreeMP -pseudo -spr 4 -mlacc 2 -slownni -gtr -nt). We rooted the tree with 10 Subtype C sequences as an outgroup, using the R package APE. The branch lengths in our tree correspond to genetic distances between the sequences, and not to time.

We also compared the results obtained from this tree with those of two trees reconstructed with RaxML (Stamatakis 2006b). These trees were reconstructed assuming a Generalized Time-Reversible evolutionary model. One assumed CAT model with 25 discrete evolutionary rate categories, the other Γ-distributed evolutionary rates. Supplementary table S1, Supplementary Material online, shows that there is generally good agreement between the results.

Heritability Estimation

To estimate the heritability of the three traits—set-point viral load, CD4+ T-cell decline and per-parasite pathogenicity—we used two approaches.

First, we applied donor–recipient regressions that are formally equivalent to parent-offspring regressions (Fraser et al. 2014) to the 196 previously identified transmission pairs. Although we do not have any information on the direction of transmission in these pairs, we expect that a regression between the trait in question will yield a good estimate of the heritability (Bachmann et al. 2017).

Second, we employed phylogenetic mixed models (Housworth et al. 2004) that are widely used to estimate the heritability from a phylogenetic tree. These methods have the advantage of being able to incorporate larger study populations and transmission relationships ranging from close pairing to distant epidemiological linkage. We used the recent implementation by Leventhal and Bonhoeffer (2016) in the R language of statistical computing (R Core Team 2013). The models that underlie this method assume trait evolution according to Brownian motion, meaning that traits drift neutrally. Brownian motion is characterized by the diffusion constant which is related to the heritability. Because we use the method on a tree, the branch lengths of which correspond to genetic distances, we make the implicit assumption that heritability increases linearly with genetic distance. Previous studies used Pagel's λ to estimate heritabilities (Alizon et al. 2010; Shirreff et al. 2013). We refrained from using Pagels λ in our main analysis because it is not as appropriate as phylogenetic mixed model for nonultrametric trees (as our phylogenetic tree; Leventhal and Bonhoeffer 2016). However, as a point of comparison with estimates of Pagels λ in previous studies, we report our estimates of this quantity in the supplementary material, Supplementary Material online.

We also applied phylogenetic mixed models based on the Ornstein–Uhlenbeck process that describe stabilizing trait selection around an optimal trait value rather than neutral drift. In addition to the diffusion constant, this model has a parameter for the trait value around which selection stabilizes the trait, θ, and another parameter describing the strength of selection, α. In macroevolutionary applications, the parameter α is often translated into a characteristic time $t_{1/2} = ln(2)/\alpha$ needed for the trait to evolve halfway back to its optimum (Hansen 1997). The heritability in POUMM is related to all three parameters of the Ornstein–Uhlenbeck process (Mitov and Stadler 2016). We used the implementation by Mitov and Stadler (2017).

Testing for Genetic Associations with Per-Parasite Pathogenicity

Previous studies identified a valine instead of an isoleucine at position 62 and 64 in the protease (I62V and I64V), and a proline instead of an alanine at position 272 in the reverse transcriptase as amino acid substitutions that could be associated with per-parasite pathogenicity (Ng et al. 2014) (see the Result section for more details). To test for associations of per-parasite pathogenicity with the substitutions I62V, I64V, and A272P, we regressed per-parasite pathogenicty, but also the other two traits, set-point viral load and CD4+ T cells decline, against categorical variables indicating the presence of the mutated amino acid at the respective position. The mutated amino acids were defined as a valine in position 62 and 64 of the protease, and a proline in position 272 of the reverse transcriptase. In these regressions, we also included cofactors such as age, sex and risk group.

An association with per-parasite pathogenicity was assessed in two ways. First, we regressed the proxy for per-parasite pathogenicity defined above—residuals from the age-adjusted tolerance curves—against the presence of the substitute amino acid. Second, we regressed the rate of change in the CD4+ T cells against the square of the logarithm to the base of 10 of the set-point viral load, including the presence of the mutated amino acid at the respective position as an interaction term. We also included the sex and age at infection as cofactors in the analysis:

$$\begin{aligned}\Delta\mathrm{CD4} = (&\alpha_0 + \eta_{62PROT.V} + \eta_{64PROT.V} \\ &+ \eta_{272RT.P} + \eta_M + c\,a)(\log_{10}V)^2 + \epsilon\end{aligned} \quad (1)$$

Here, ΔCD4 denotes the rate of change in the CD4+ T-cell level per microliter blood per day. The parameter α_0 describes the relationship between set-point viral load and CD4+ T cells decline for females with age zero that are infected with a virus that does not express any of the mutant amino acids we consider. The remaining parameters describe the offset that can be attributed to the various cofactors: $\eta_{62PROT.V}$, $\eta_{64PROT.V}$, and $\eta_{272RT.P}$ describe the potential change in the relationship between set-point viral load and CD4+ T-cell decline due to each considered amino acid substitution, and η_M quantifies the change due to being male. *ca* is the offset to α_0 in an individual of age *a*. This procedure follows the approach we have adopted previously to test for the association of host factors with disease tolerance (Regoes et al. 2014).

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

We are grateful to Oliver Laeyendecker for making us aware of the amino acid polymorphisms that could be associated with per-parasite pathogenicity.

We thank the patients who participate in the Swiss HIV Cohort Study (SHCS); the physicians and study nurses, for excellent patient care; the resistance laboratories, for high-quality genotypic drug resistance testing; SmartGene (Zug, Switzerland), for technical support; Brigitte Remy, RN, Martin Rickenbach, MD, Franziska Schöni-Affolter, MD, and Yannick Vallet, MSc, from the SHCS Data Center (Lausanne, Switzerland), for data management; and Danièle Perraudin and Mirjam Minichiello for administrative assistance.

R.R.R. acknowledges the financial support of the Swiss National Science Foundation (grant number: 31003A_149769). R.D.K. and A.M. have been supported by the Swiss National Science Foundation (Grant number: BSSGI0_155851). V.M. was supported by the grant GINOP-2.3.2-15-2016-00057 ("Az evolúció fényében: elvek és megoldások"). This study has been performed within the framework of the Swiss HIV Cohort Study, supported by the Swiss National Science Foundation (grant number 33CS30_148522) and was further supported by the SHCS research foundation. The Swiss HIV Drug Resistance database is supported by SNF project 320030_159868 to H.F.G., the Yvonne-Jacob Foundation; Gilead, Switzerland (one unrestricted grant to the SHCS Research Foundation and one unrestricted grant to H.F.G.); and the University of Zurich's Clinical Research Priority Program (Viral Infectious Diseases: Zurich Primary HIV Infection Study; to H.F.G.).

The data are gathered by the 5 Swiss University Hospitals, 2 Cantonal Hospitals, 15 affiliated hospitals, and 36 private physicians (listed in http://www.shcs.ch/31-health-care-providers; last accessed September 24, 2017). The members of the Swiss HIV Cohort Study are: Aubert V, Barth J, Battegay M, Bernasconi E, Böni J, Bucher HC, Burton-Jeangros C, Calmy A, Cavassini M, Egger M, Elzi L, Fehr J, Fellay J, Furrer H (Chairman of the Clinical and Laboratory Committee), Fux CA, Gorgievski M, Günthard H (President of the SHCS), Haerry D (deputy of "Positive Council"), Hasse B, Hirsch HH, Hösli I, Kahlert C, Kaiser L, Keiser O, Klimkait T, Kouyos R, Kovari H, Ledergerber B, Martinetti G, Martinez de Tejada B, Metzner K, Müller N, Nadal D, Pantaleo G, Rauch A (Chairman of the Scientific Board), Regenass S, Rickenbach M (Head of Data Center), Rudin C (Chairman of the Mother & Child Substudy), Schöni-Affolter F, Schmid P, Schultze D, Schüpbach J, Speck R, Staehelin C, Tarr P, Telenti A, Trkola A, Vernazza P, Weber R, Yerly S.

References

Alizon S, von Wyl V, Stadler T, Kouyos RD, Yerly S, Hirschel B, Böni J, Shah C, Klimkait T, Furrer H, et al. 2010. Phylogenetic approach reveals that virus genotype largely determines HIV set-point viral load. *PLoS Pathog*. 6(9):e1001123.

Arnaout RA, Lloyd AL, O'Brien TR, Goedert JJ, Leonard JM, Nowak MA. 1999. A simple relationship between viral load and survival time in HIV-1 infection. *Proc Natl Acad Sci U S A*. 96(20):11549–11553.

Ayres JS, Schneider DS. 2012. Tolerance of infections. *Annu Rev Immunol*. 30:271–294.

Bachmann N, Turk T, Kadelka C, Marzel A, Shilaih M, Böni J, Aubert V, Klimkait T, Leventhal GE, Günthard HF, et al. 2017. Parent-offspring regression to estimate the heritability of an HIV-1 trait in a realistic setup. *Retrovirology* 14(1):33.

Baeten JM, Chohan B, Lavreys L, Chohan V, McClelland RS, Certain L, Mandaliya K, Jaoko W, Overbaugh J. 2007. HIV-1 subtype D infection is associated with faster disease progression than subtype A in spite of similar plasma HIV-1 loads. *J Infect Dis*. 195(8):1177–1180.

Barbour JD, Hecht FM, Wrin T, Segal MR, Ramstead CA, Liegler TJ, Busch MP, Petropoulos CJ, Hellmann NS, Kahn JO, et al. 2004. Higher CD4+ T cell counts associated with low viral pol replication capacity among treatment-naive adults in early HIV-1 infection. *J Infect Dis*. 190(2):251–256.

Bartha I, Carlson JM, Brumme CJ, McLaren PJ, Brumme ZL, John M, Haas DW, Martinez-Picado J, Dalmau J, López-Galíndez C, et al. 2013. A genome-to-genome analysis of associations between human genetic variation, HIV-1 sequence diversity, and viral control. *Elife* 2:e01123.

Bartha I, McLaren PJ, Brumme C, Harrigan R, Telenti A, Fellay J. 2017. Estimating the respective contributions of human and viral genetic variation to HIV control. *PLoS Comput Biol*. 13(2):e1005339.

Blanquart F, Wymant C, Cornelissen M, Gall A, Bakker M, Bezemer D, Hall M, Hillebregt M, Ong SH, Albert J, on behalf of the BEEHIVE collaboration, et al. 2017. Viral genetic variation accounts for a third of variability in HIV-1 set-point viral load in Europe. *PLoS Biol*. 15(6):e2001855.

Boots M. 2008. Fight or learn to live with the consequences? *Trends Ecol Evol (Amst)*. 23(5):248–250.

Boots M, Best A, Miller MR, White A. 2009. The role of ecological feedbacks in the evolution of host defence: what does theory tell us? *Philos Trans R Soc Lond B, Biol Sci*. 364(1513):27–36.

Caldwell RM, Schafer JF, Compton LE, Patterson FL. 1958. Tolerance to cereal leaf rusts. *Science* 128(3326):714–715.

Carlson JM, Schaefer M, Monaco DC, Batorsky R, Claiborne DT, Prince J, Deymier MJ, Ende ZS, Klatt NR, DeZiel CE, et al. 2014. HIV transmission. Selection bias at the heterosexual HIV-1 transmission bottleneck. *Science* 345(6193):1254031.

Carlson JM, Du VY, Pfeifer N, Bansal A, Tan VYF, Power K, Brumme CJ, Kreimer A, DeZiel CE, Fusi N, et al. 2016. Impact of pre-adapted HIV transmission. *Nat Med*. 22(6):606–613.

Chahroudi A, Bosinger SE, Vanderford TH, Paiardini M, Silvestri G. 2012. Natural SIV hosts: showing AIDS the door. *Science* 335(6073):1188–1193.

Chakrabarti LA. 2004. The paradox of simian immunodeficiency virus infection in sooty mangabeys: active viral replication without disease progression. *Front Biosci*. 9(1–3):521–539.

Daar ES, Kesler KL, Petropoulos CJ, Huang W, Bates M, Lail AE, Coakley EP, Gomperts ED, Donfield SM. 2007. Baseline HIV type 1 coreceptor tropism predicts disease progression. *Clin Infect Dis*. 45(5):643–649.

Deeks SG, Coleman RL, White R, Pachl C, Schambelan M, Chernoff DN, Feinberg MB. 1997. Variance of plasma human immunodeficiency virus type 1 RNA levels measured by branched DNA within and between days. *J Infect Dis*. 176(2):514–517.

Edgar RC. 2004. Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. 32(5):1792.

Fellay J, Shianna KV, Ge D, Colombo S, Ledergerber B, Weale M, Zhang K, Gumbs C, Castagna A, Cossarizza A, et al. 2007. A whole-genome association study of major determinants for host control of HIV-1. *Science* 317(5840):944–947.

Fraser C, Lythgoe K, Leventhal GE, Shirreff G, Hollingsworth TD, Alizon S, Bonhoeffer S. 2014. Virulence and pathogenesis of HIV-1 infection: an evolutionary perspective. *Science* 343(6177):1243727.

Goetz MB, Leduc R, Wyman N, Kostman JR, Labriola AM, Lie Y, Weidler J, Coakley E, Bates M, Luskin-Hawk R. 2010. HIV replication capacity is an independent predictor of disease progression in persons with untreated chronic HIV infection. *J Acquir Immune Defic Syndr*. 53(4):472–479.

Goulder PJR, Watkins DI. 2008. Impact of MHC class I diversity on immune control of immunodeficiency virus replication. *Nat Rev Immunol*. 8(8):619–630.

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

Hodcroft E, Hadfield JD, Fearnhill E, Phillips A, Dunn D, O'Shea S, Pillay D, Leigh Brown AJ, Worobey M. 2014. The contribution of viral genotype to plasma viral set-point in HIV infection. *PLoS Pathog*. 10(5):e1004112.

Hollingsworth TD, Laeyendecker O, Shirreff G, Donnelly CA, Serwadda D, Wawer MJ, Kiwanuka N, Nalugoda F, Collinson-Streng A, Ssempijja V, et al. 2010. HIV-1 transmitting couples have similar viral load set-points in Rakai, Uganda. *PLoS Pathog*. 6(5):e1000876.

Housworth EA, Martins EP, Lynch M. 2004. The phylogenetic mixed model. *Am Nat*. 163(1):84–96.

Kirchhoff F. 2009. Is the high virulence of HIV-1 an unfortunate coincidence of primate lentiviral evolution? *Nat Rev Microbiol*. 7(6):467–476.

Kouyos RD, Rauch A, Böni J, Yerly S, Shah C, Aubert V, Klimkait T, Kovari H, Calmy A, Cavassini M, et al. 2014. Clustering of HCV coinfections on HIV phylogeny indicates domestic and sexual transmission of HCV. *Int J Epidemiol*. 43(3):887–896.

Kozlov AM, Aberer AJ, Stamatakis A. 2015. Examl version 3: a tool for phylogenomic analyses on supercomputers. *Bioinformatics* 31(15):2577.

Leventhal GE, Bonhoeffer S. 2016. Potential pitfalls in estimating viral load heritability. *Trends Microbiol*. 24(9):687–698.

Little TJ, Shuker DM, Colegrave N, Day T, Graham AL. 2010. The coevolution of virulence: tolerance in perspective. *PLoS Pathog*. 6(9):e1001006.

Medzhitov R, Schneider DS, Soares MP. 2012. Disease tolerance as a defense strategy. *Science* 335(6071):936–941.

Mellors JW, Rinaldo CR, Gupta P, White RM, Todd JA, Kingsley LA. 1996. Prognosis in HIV-1 infection predicted by the quantity of virus in plasma. *Science* 272(5265):1167–1170.

Mitov V, Stadler T. 2016. The heritability of pathogen traits – definitions and estimators. *bioRxiv* 058503.

Mitov V, Stadler T. 2017. POUMM: An R-package for Bayesian Inference of Phylogenetic Heritability. *bioRxiv* 115089.

Müller V, Fraser C, Herbeck JT. 2011. A strong case for viral genetic factors in HIV virulence. *Viruses* 3(3):204–216.

Ng OT, Laeyendecker O, Redd AD, Munshaw S, Grabowski MK, Paquet AC, Evans MC, Haddad M, Huang W, Robb ML, et al. 2014. HIV type 1 polymerase gene polymorphisms are associated with phenotypic differences in replication capacity and disease progression. *J Infect Dis*. 209(1):66–73.

Pantazis N, Porter K, Costagliola D, De Luca A, Ghosn J, Guiguet M, Johnson AM, Kelleher AD, Morrison C, Thiebaut R, et al. 2014. Temporal trends in prognostic markers of HIV-1 virulence and transmissibility: an observational cohort study. *Lancet HIV*. 1(3):e119–e126.

Phillips RE, Rowland-Jones S, Nixon DF, Gotch FM, Edwards JP, Ogunlesi AO, Elvin JG, Rothbard JA, Bangham CRM, Rizza CR et al. 1991. Human immunodeficiency virus genetic variation that can escape cytotoxic t cell recognition. *Nature* 354:453–459.

Price MN, Dehal PS, Arkin AP, Poon AFY. 2010. FastTree 2–approximately maximum-likelihood trees for large alignments. *PLoS ONE*. 5(3):e9490.

R Core Team 2013. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Råberg L. 2014. How to live with the enemy: understanding tolerance to parasites. *PLoS Biol*. 12(11):e1001989.

Råberg L, Stjernman M. 2012. The evolutionary ecology of infectious disease virulence. In: Damas G, Nelson, R, editors. Ecoimmunology. New York: Oxford University Press, p. 548578.

Råberg L, Sim D, Read AF. 2007. Disentangling genetic variation for resistance and tolerance to infectious diseases in animals. *Science* 318(5851):812–814.

Råberg L, Graham AL, Read AF. 2009. Decomposing health: tolerance and resistance to parasites in animals. *Philos Trans R Soc Lond. B, Biol Sci*. 364(1513):37–49.

Raboud JM, Montaner JSG, Conway B, Haley L, Sherlock C, O'Shaughnessy MV, Schechter MT. 1996. Variation in plasma RNA levels, CD4 cell counts, and p24 antigen levels in clinically stable men with human immunodeficiency virus infection. *J Infect Dis*. 174(1):191–194.

Read AF, Graham AL, Råberg L. 2008. Animal defenses against infectious agents: is damage control more important than pathogen control. *PLoS Biol*. 6(12):e4.

Regoes RR, McLaren PJ, Battegay M, Bernasconi E, Calmy A, Günthard HF, Hoffmann M, Rauch A, Telenti A, Fellay J. 2014. Disentangling human tolerance and resistance against HIV. *PLoS Biol*. 12(9):e1001951.

Rodriguez B, Sethi AK, Cheruvu VK, Mackay W, Bosch RJ, Kitahata M, Boswell SL, Mathews WC, Bangsberg DR, Martin J, et al. 2006. Predictive value of plasma HIV RNA level on rate of CD4 T-cell decline in untreated HIV infection. *JAMA* 296(12):1498–1506.

Schafer J. 1971. Tolerance to plant disease. *Annu Rev Phytopathol*. 9(1):235–252.

Schneider DS, Ayres JS. 2008. Two ways to survive infection: what resistance and tolerance can teach us about treating infectious diseases. *Nat Rev Immunol*. 8(11):889–895.

Schoeni-Affolter F, Ledergerber B, Rickenbach M, Rudin C, Günthard HF, Telenti A, Furrer H, Yerly S, Francioli P. 2010. Cohort profile: the Swiss HIV Cohort study. *Int J Epidemiol*. 39(5):1179–1189.

Shirreff G, Alizon S, Cori A, Günthard HF, Laeyendecker O, van Sighem A, Bezemer D, Fraser C. 2013. How effectively can HIV phylogenies be used to measure heritability? *Evol Med Public Health*. 2013(1):209–224.

Stamatakis A. 2006a. Phylogenetic models of rate heterogeneity: a high performance computing perspective. In: *Proceedings 20th IEEE International Parallel & Distributed Processing Symposium*, IEEE, p. 8.

Stamatakis A. 2006b. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22(21):2688.

van Dorp CH, van Boven M, de Boer RJ, Regoes RR. 2014. Immunoepidemiological modeling of HIV-1 predicts high heritability of the set-point virus load, while selection for CTL escape dominates virulence evolution. *PLoS Comput Biol*. 10(12):e1003899.

von Wyl V, Yerly S, Böni J, Bürgisser P, Klimkait T, Battegay M, Furrer H, Telenti A, Hirschel B, Vernazza PL, et al. 2007. Emergence of HIV-1 drug resistance in previously untreated patients initiating combination antiretroviral treatment: a comparison of different regimen types. *Arch Intern Med*. 167(16):1782–1790.

Yang W-L, Kouyos R, Scherrer AU, Böni J, Shah C, Yerly S, Klimkait T, Aubert V, Furrer H, Battegay M, et al. 2015. Assessing the paradox between transmitted and acquired HIV type 1 drug resistance mutations in the Swiss HIV Cohort Study from 1998 to 2012. *J Infect Dis*. 212(1):28–38.

Zheng Y-H, Jeang K-T, Tokunaga K. 2012. Host restriction factors in retroviral infection: promises in virus–host interaction. *Retrovirology* 9(1):112.

APPENDIX

Table S1: Comparison of the heritability estimates using different tree reconstruction algorithms. For all traits, POUMM fits the data significantly better than PMM.

Method	Tree reconstruction	ΔCD4 (unadjusted)	ΔCD4 (adjusted)	spVL (unadjusted)	spVL (adjusted)	ppp
PMM (ML)	FastTree	25% (9%–40%)	24% (7%–39%)	12% (2%–28%)	8% (0%–26%)	22% (5%–39%)
	RaxML-CAT	24% (9%–39%)	0% (0%–10%)	13% (3%–29%)	0% (0%–12%)	23% (7%–39%)
	RaxML-Gamma	24% (9%–39%)	0% (0%–3%)	13% (3%–29%)	11% (2%–25%)	23% (6%–39%)
PMM (MCMC)	FastTree	24% (12%–36%)	23% (9%–36%)	13% (4%–23%)	9% (0%–19%)	21% (9%–32%)
	RaxML-CAT	24% (13%–34%)	25% (14%–34%)	14% (5%–24%)	9% (0%–20%)	24% (13%–35%)
	RaxML-Gamma	24% (13%–35%)	24% (14%–36%)	14% (5%–25%)	11% (1%–22%)	22% (6%–33%)
POUMM	FastTree	17% (6%–29%)	17% (5%–30%)	26% (8%–43%)	29% (12%–46%)	17% (4%–29%)
	RaxML-CAT	20% (8%–30%)	21% (9%–32%)	33% (19%–48%)	38% (22%–52%)	19% (7%–31%)
	RaxML-Gamma	19% (8%–30%)	20% (8%–33%)	30% (15%–46%)	34% (18%–50%)	19% (6%–31%)

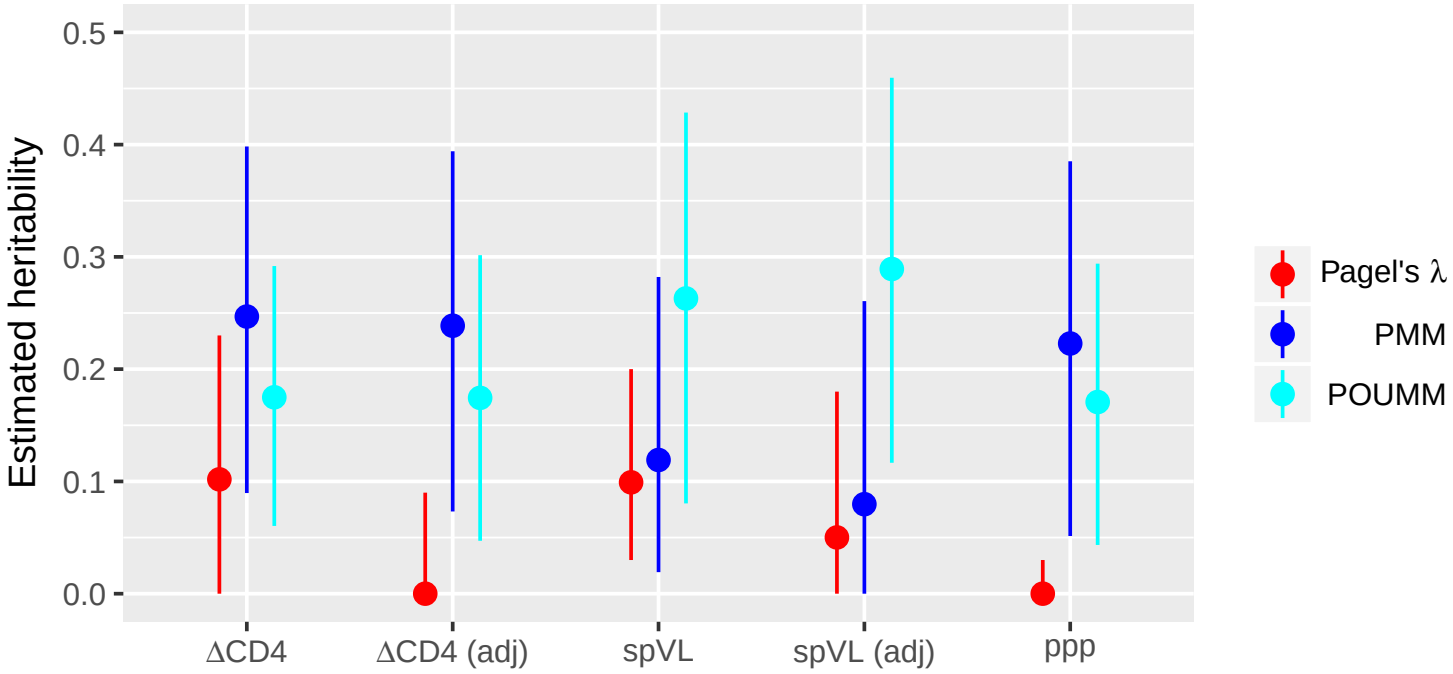


Figure S1: Heritability estimates using using Pagel's λ in comparison to those using PMM and POUMM. The points are the point estimates, the vertical lines show the 95% confidence intervals. We show these estimates only for comparison with older papers that used Pagel's λ on unadjusted traits. Because of the non-ultrametric nature of our phylogenetic tree Pagel's λ is not appropriate to use.

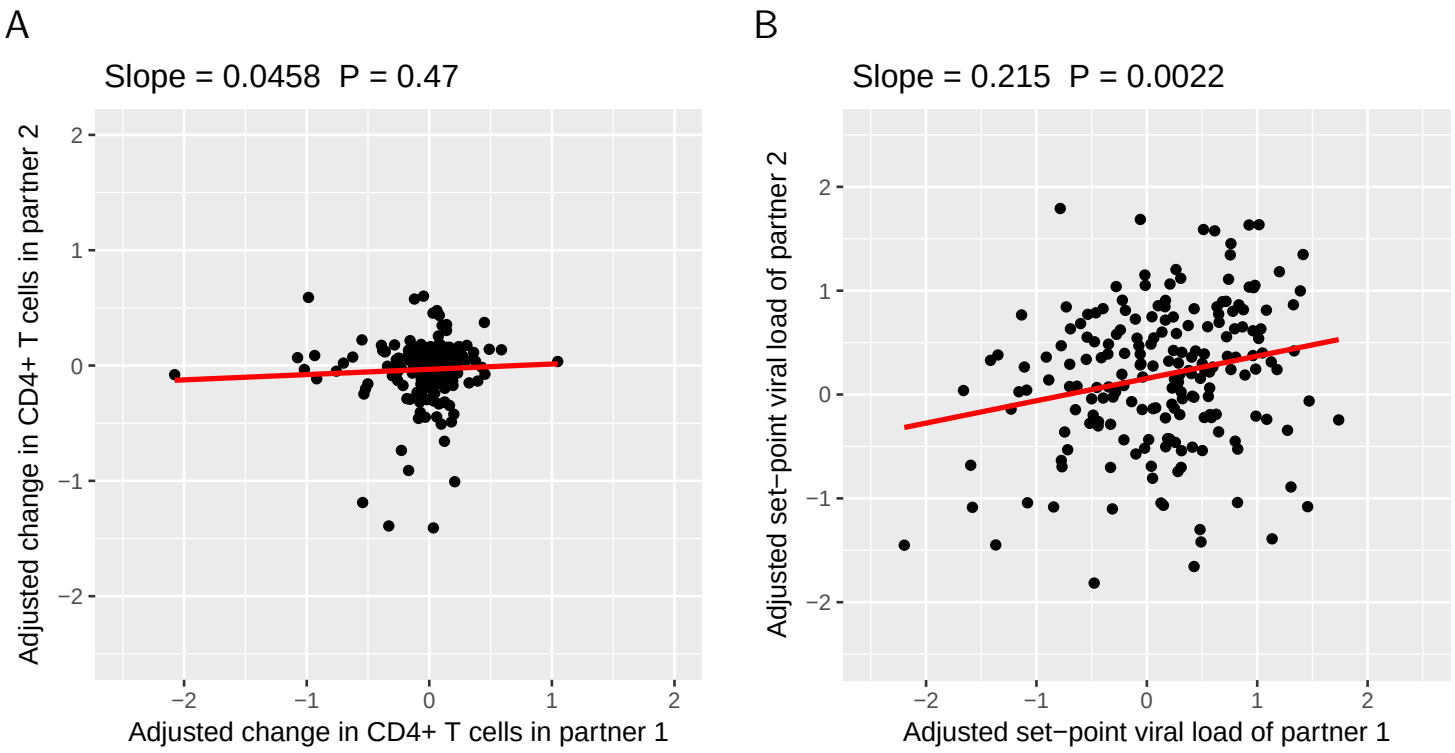


Figure S2: Heritability estimates from donor-recipient regressions on adjusted CD4+ T cell decline and set-point viral loads. For the definition of the adjusted trait values see the Methods and Materials. (See also Fig 3.)

Part III

MANUSCRIPTS

5

PARALLEL LIKELIHOOD CALCULATION FOR GAUSSIAN PHYLOGENETIC MODELS

In major revision for a publication as

Venelin Mitov and Tanja Stadler (2018). Fast Bayesian Inference of Gaussian Phylogenetic Models Using Parallel Likelihood Calculation. *Methods in Ecology and Evolution*.

In this article I present the C++ library SPLITT for Serial and Parallel Lineage Traversal of Trees. The origin of this tool has been the POUMM R-package used to estimate the heritability of HIV set-point viral load in the previous two chapters. SPLITT was designed as a low-level library that deals with the technical difficulties of parallel or serial tree traversal operations while exposing a handy user interface to higher-level tools. This library is now used as a back-end by the POUMM R-package and the PCMBase R-package, which will be introduced in the next chapter 6. In the Appendix of this chapter, I describe a quadratic polynomial representation of the likelihood of a single-trait POUMM model. This theoretical development inspired the next Chapter, where I will describe a generalization of this approach to a larger family of models.

ABSTRACT

1. Gaussian phylogenetic models have been used to model trait evolution, to measure the heritability of traits, to test selection versus neutral hypotheses, to estimate optimal trait-values, and to quantify rates of adaptation. Despite the existence of linear algorithms for calculating the likelihood of Gaussian models, Bayesian inference on large trees keeps being a time-intensive task.

2. Speeding-up Bayesian inference is an active field in applied Statistics with numerous recent developments, ranging from Metropolis sampling with adaptive proposal to parallel MCMC sampling of conditionally independent parameters and parallel likelihood calculation assuming independent traits or using parallel linear algebra libraries. For different reasons, few of these techniques apply to Gaussian phylogenetic models. Here, we introduce parallel likelihood calculation based on parallel tree traversal. Parallel tree traversal has been used in Computer science to automate the scheduling of dependant tasks, but has, to our knowledge, not been applied in phylogenetic modeling.

3. We implement several parallel algorithms in the form of a C++ library for Serial and Parallel LIneage Traversal of Trees (SPLITT). Using univariate and multivariate versions of a phylogenetic Ornstein-Uhlenbeck mixed model (POUMM), we run benchmarks on up to 24 CPU cores, reporting up to an order of magnitude parallel speed-up on simulated balanced and unbalanced trees of up to 100,000 tips with up to 16 traits. Noticing that the parallel speed-up depends on multiple factors, the SPLITT library is capable to automatically select the fastest traversal strategy for a given hardware, tree-topology and data. Combining SPLITT likelihood calculation with adaptive Metropolis sampling on real data, we show that the time for Bayesian POUMM inference on a tree of 10,000 tips can be reduced from several days to minutes.

4. We conclude that parallel tree traversal effectively accelerates the likelihood calculation of Gaussian phylogenetic models. For fastest Bayesian inference, we recommend combining this technique with adaptive Metropolis sampling. Beyond Gaussian models, the parallel tree traversal can be applied to numerous other models, including discrete trait and birth-death models. Currently, SPLITT supports multi-core shared memory architectures, but can be extended to distributed memory architectures as well as graphical processing units.
Keywords: post-order traversal, pre-order traversal, discrete character, continuous trait, phylogenetic comparative models, Brownian motion

5.1 INTRODUCTION

The past decades have seen active development of phylogenetic comparative models (PCMs) of trait evolution, progressing from null neutral models, such as single-trait Brownian mo-

tion (BM), to complex multi-trait models incorporating selection, interaction between trait values and diversification, and co-evolution of traits (Manceau, Lambert, and Morlon, 2016; O'Meara, 2012). Recent works have shown that, for a broad family of PCMs, the likelihood of an observed tree and data conditioned on the model parameters can be computed in time proportional to the size of the tree (FitzJohn, 2012; Goolsby, Bruggeman, and Ané, 2016; Ho and Ané, 2014a; Manceau, Lambert, and Morlon, 2016). This family includes Gaussian models like Brownian motion (BM) and Ornstein-Uhlenbeck (OU) phylogenetic models as well as some non-Gaussian models like phylogenetic logistic regression (Ho and Ané, 2014a; Ives and Garland, 2010; Paradis and Claude, 2002). All of these likelihood calculation techniques rely on post-order tree traversal also known as *"pruning"* (Felsenstein, 1973, 1981; Felsenstein, 1983). For moderate numbers of traits, combining pruning algorithms for likelihood calculation with gradient-based optimization (Boyd and Vandenberghe, 2004) enables maximum likelihood model inference within seconds on contemporary computers, even for phylogenies of many thousands of tips (Ho and Ané, 2014a). Despite its simple interpretation and several useful statistical properties, the maximum likelihood estimator (MLE) has often been criticised for being a point estimator, prone to be a local optimum and uninformative about the likelihood surface.

As an elegant alternative, Bayesian approaches such as Markov Chain Monte Carlo (MCMC) allow incorporating prior knowledge in the model inference and provide posterior samples and high posterior density (HPD) intervals for the model parameters (FitzJohn, 2012; Slater, Harmon, and Alfaro, 2012). In contrast with ML inference, though, Bayesian inference methods require many orders of magnitude more likelihood evaluations. This presents a bottleneck in Bayesian analysis, in particular, for complex models of many unknown parameters or when faced with large phylogenies of many thousands of tips, such as transmission trees from large-scale epidemiological studies, e.g. Alizon et al. (2010), Bachmann et al. (2017), Bertels et al. (2017), Blanquart et al. (2017), Hodcroft et al. (2014), Mitov and Stadler (2018), and Shirreff et al. (2013). While big data should provide the needed statistical power to fit a complex model, the time needed to perform a full scale Bayesian fit often limits the choice to a faster but less informative ML-inference, or a Bayesian inference of a simplified model.

Speeding-up Bayesian inference is an active topic in applied Statistics with recent advances that can be classified in several groups. One group of methods are adaptive variants of the random walk Metropolis (RWM) algorithm (Metropolis et al., 1953) that aim to decrease the number of MCMC iterations by performing "on-the-fly" changes of the jump distribution, based on what has been "learned" about the parameter space from past iterations (Haario, Saksman, and Tamminen, 2001; Vihola, 2012). A major advantage of these methods is that they are generic with respect to the models and can be implemented as general purpose Metropolis samplers (e.g. `adaptMCMC` (Scheidegger, 2012)). A second group are "pre-fetching" methods which modify the Metropolis-Hastings algorithm so that it speculatively executes sequences of individual likelihood calls in parallel, "hoping" that these sequences tend to match the actual accepted states of the MCMC (Angelino et al., 2014; Brockwell, 2006). Another possibility to use multiple processor power, which could potentially be combined with the above methods is to delegate the parallelization problem to a low level linear algebra library, e.g. OpenBLAS (Wang et al., 2013).

A separate body of work, to which this work counts, is the ensemble of model-specific approaches that parallelize the likelihood calculation by using specific features of the likelihood function. These include factorizations of the likelihood into a product of components associated with conditionally independent subsets of the model parameters (Goudie et al., 2017; Whiley and Wilson, 2004) or the observed variables (Ayres et al., 2012). Often, this factorization relies on strong model assumptions, such as a hierarchical structure of the model

parameters or independence of the observed variables. A common approach used in software packages like BEAST (Bouckaert et al., 2014; Drummond et al., 2012) is to combine the factorization with caching and reusing of some of the previously calculated likelihood components in consecutive MCMC iterations, as long as these are not affected by the proposed jump in parameter space.

For a Gaussian phylogenetic model, though, the likelihood cannot be factorized across parameter groups, trait independence is acceptable only as a null hypothesis and, with a moderate number of traits and pruning-wise likelihood calculation, parallelizing algebraic operations (on low-dimensional vectors and matrices) is inefficient. Hence, we explore the parallelization of the likelihood calculation at the level of traversing the phylogenetic tree, that is, the pruning. Parallel tree traversal has been studied in Computer science, mostly for the purposes of parallel tree contraction (Reif, 1989) and automated task scheduling (Qamnieh, 2015). Capitalizing on the same ideas, we developed SPLITT: a shared-memory C++ library for Serial and Parallel Lineage Traversal of Trees. While we focus on Gaussian phylogenetic models as the main application of the library, we designed the SPLITT programming interface to be generic with respect to the node-visiting operation, hoping that the library could potentially find use in different models, including birth-death population models and discrete trait models. We tested SPLITT on large trees (up to N=100,000) and on different topologies, including balanced and highly unbalanced trees. These tests proved a nice property of the parallel pruning algorithm, namely the fact that its parallel efficiency increases with the tree size as well as the complexity of the node-visiting operation. Thus, for large trees and complex models, the parallel speed-up is limited either by the number of available processors or by another limited resource such as the memory bandwidth.

5.2 MATERIALS AND METHODS

5.2.1 *Setup*

Through the rest of the article we will use the following notation. Given is a rooted phylogenetic tree $\mathcal{T}$ with a total of M nodes, including $N < M$ tips denoted $1, ..., N$, $M - N - 1$ internal nodes denoted $N + 1, ..., M - 1$, and a root node denoted M (Fig. 5.1). Without restrictions on the tree topology, non-ultrametric trees (i.e. tips have different heights) and polytomies (i.e. nodes with any finite number of descendants) are accepted. We denote by $\mathcal{T}_i$ the subtree rooted at node i. For any tip or internal node i, we denote its parent node by $Parent(i)$. For any node j, we denote by $Desc(j)$ the set of its direct descendants ($Desc(j) = \phi$ if j denotes a tip). Furthermore, for any $i \in Desc(j)$, we denote by t_i the length of the branch leading to i. Associated with each node i there is an input data in the form of a single or multivariate categorical or numerical value denoted z_i. For tips, z_i can be partially unobserved (having NA entries), while for internal nodes or the root it can also be fully absent (NULL). We denote by $\mathbf{z}_i$ the sub-vector of input data for the nodes in $\mathcal{T}_i$. Associated with each node, i, there is a vector of model parameters, Θ_i. We use bold style $\mathbf{t}$, $\mathbf{z}$ and $\mathbf{\Theta}$ when denoting the vectors of all branch lengths, input data and parameters.

5.2.2 *A general framework for parallel tree traversal*

Let $F_{\mathcal{T}}(\mathbf{t}, \mathbf{z}, \mathbf{\Theta})$ be a function of the branch lengths, the input data and the parameters. A post-order tree traversal algorithm can be used to calculate $F_{\mathcal{T}}$ if, for all subtrees $\mathcal{T}_j$ of $\mathcal{T}$, there exist functions $S_j(\mathbf{t}, \mathbf{z}, \mathbf{\Theta})$, hereby called "states", satisfying the following rules:

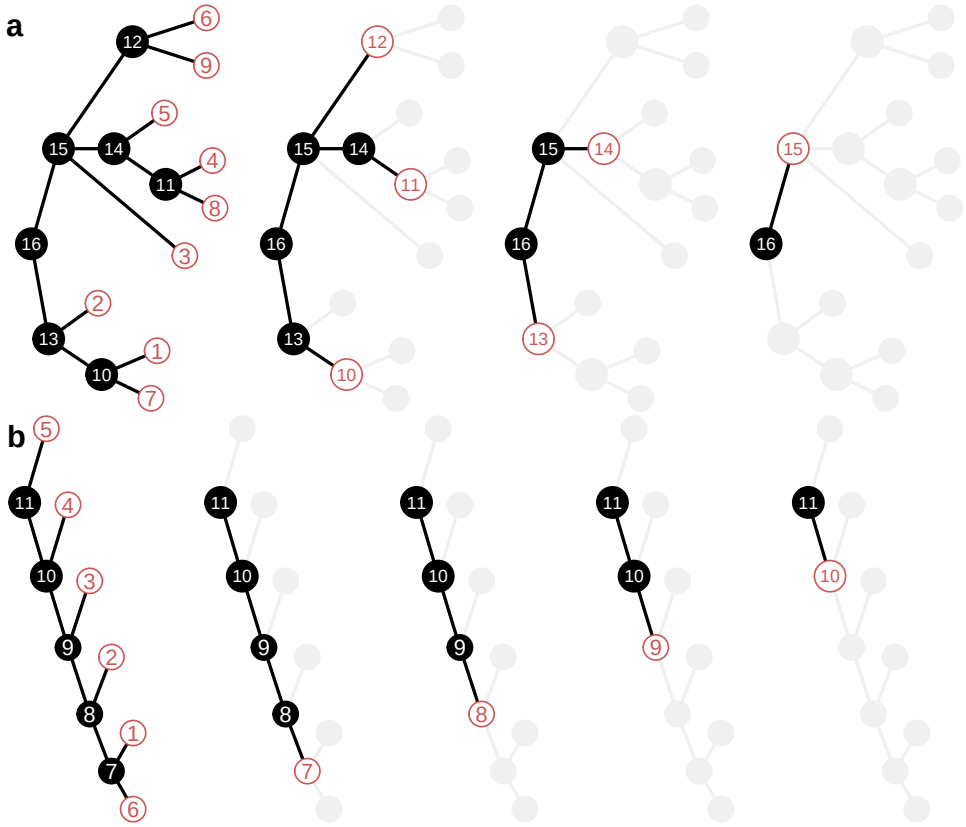


Figure 5.1: **Parallel pruning.** The trees from left to right depict generations of nodes that can be processed in parallel. The processing of a node consists in calculating its state based on the input data, the branch lengths and the states of the node's direct descendants (eq. 5.1). black: nodes having one or more non-processed descendants; red: nodes ready to be processed; grey: nodes processed in a previous generation. a) a balanced tree; b) a ladder.

(1) $F_{\mathcal{T}}(\mathbf{t}, \mathbf{z}, \mathbf{\Theta})$ can be calculated from $S_M(\mathbf{t}, \mathbf{z}, \mathbf{\Theta})$;
(2) For each node $j \in \{1, ..., M\}$, there exists a (recursive) relationship R_j between S_j and the set of states at j's descendants, such that:

$$S_j(\mathbf{t}, \mathbf{z}, \mathbf{\Theta}) = R_j\Big(\big\{S_i(\mathbf{t}, \mathbf{z}, \mathbf{\Theta}) : i \in Desc(j)\big\}, \mathbf{t}, \mathbf{z}, \mathbf{\Theta}\Big). \tag{5.1}$$

We note that analogical terms can be defined for pre-order tree traversal. In this case the target functions are values $Z_{\mathcal{T},j}(\mathbf{t}, \mathbf{z}, \mathbf{\Theta})$ corresponding to the nodes $j \in \{1, ..., M\}$, and rule (2) is updated to:

(2') Z_M can be calculated from the input data. For each node $j \in \{1, ..., M-1\}$, there exists a (recursive) relationship R'_j between Z_j and $Z_{Parent(j)}$, such that:

$$Z_j(\mathbf{t}, \mathbf{z}, \mathbf{\Theta}) = R'_j\big(Z_{Parent(j)}(\mathbf{t}, \mathbf{z}, \mathbf{\Theta}), \mathbf{t}, \mathbf{z}, \mathbf{\Theta}\big). \tag{5.2}$$

The states, i.e. the values of the functions S_j and Z_j, may be deterministic or stochastic functions of the input tree and data. They can be real numbers, vectors, matrices or higher order combinations thereof. In Supplementary information, we provide example usages of the parallel traversal framework. In each of these examples, we solve a particular problem, such as calculating the likelihood of a continuous time Markov model for a categorical or a continuous trait. In terms of the framework, the task boils down to formulating the node states $S_j(\mathbf{t}, \mathbf{z}, \mathbf{\Theta})$ and the recursive functions R_j satisfying rules (1) and (2).

For the rest of the article, we focus on parallel post-order tree traversal or "pruning", noting that the algorithms for parallel pre-order traversal are simple analogies. The SPLITT library implements both traversal types.

Rule (2) ensures that calculating the state of a node j can be done independently from the calculation of any other node k, provided that neither j is an ancestor of k, nor k is an ancestor of j. Based on this observation, we describe two alternative parallel algorithms for calculating the root state S_M, noting that similar formulations of these algorithms can be found in the Computer science literature (Qamnieh, 2015; Reif, 1989).

5.2.2.1 *Queue-based parallel pruning*

It is possible to parallelize the computation of the states S_j across multiple computing threads using a first-in-first-out list (queue) of the nodes in the tree (algorithm 5.1). Initially, the queue is filled with all tips in the tree and a counter with the number of direct descendants is set for each internal or root node. Then, each thread takes a node i from the front of the queue, calculates its state and decrements the counter of $Parent(i)$. If the counter of $Parent(i)$ has become zero, $Parent(i)$ is added to the queue, so that it will be processed as soon as a free thread picks it from the queue. Assuming an unlimited number of threads and a negligible cost of the queue- and the counter- operations, algorithm 5.1 guarantees that a node will be processed immediately after all of its direct descendants have been processed. Thus, in theory, algorithm 5.1 maximizes the parallel execution. However, an implementation of the atomic operations on the queue and the counters would have to rely on a thread synchonization mechanism such as a mutex, which can be slow on some systems. Thus, a decent parallelization speed-up would only be possible if the overall cost of synchronization is insignificant compared to the functions R_j.

Algorithm 5.1 : Queue-based parallel pruning

Input : $\mathcal{T}$, $\mathbf{t}$, $\mathbf{z}$, $\boldsymbol{\Theta}$
Output : $S_M(\mathbf{t}, \mathbf{z}, \boldsymbol{\Theta})$
/* a vector of M states */
1 $State \longleftarrow [...]_M$;
/* a vector of the numbers of remaining descendants for each node */
2 $NumDesc \longleftarrow \left[|Desc(i)| : i \in \{1, ..., M\}\right]$;
/* initiate $Queue$ with all tips: */
3 $Queue \longleftarrow [1, ..., N]$;
4 **begin** Parallel block
5 **while** *(TRUE)* **do**
/* if $Queue$ is empty, thread waits. */
6 $j \longleftarrow$PopFirst($Queue$);
7 $State[j] \longleftarrow R_j\left(\left\{State[i] : i \in Desc(j)\right\}, \mathbf{t}, \mathbf{z}, \boldsymbol{\Theta}\right)$;
8 **if** $(j < M)$ **then**
/* the root has not been processed yet. */
9 $NumDesc[Parent(j)] \longleftarrow NumDesc[Parent(j)] - 1$;
10 **if** $(NumDesc[Parent(j)] == 0)$ **then**
/* If $Queue$ is currently empty a waiting thread will be notified. */
11 AddLast($Queue$, $Parent(j)$);
12 **else**
/* the root has been processed. */
/* Notify waiting threads by adding a stopping node-id to $Queue$. */
13 AddLast($Queue$, $M + 1$);
/* All work done, exit the loop. */
14 break;
15 **return** $State[M]$;

5.2.2.2 *Range-based parallel pruning*

We consider an alternative of algorithm 5.1 minimizing the synchronization overhead. This approach consists in splitting the tree into "generations" of nodes, such that nodes within a generation can be processed in random order and in parallel, but only if all generations containing descendants of these nodes have already been processed (fig. 5.1). A "master" thread is responsible for launching a team of "worker" threads on each generation, starting from a generation of all tips, then taking their parents, and so on until reaching the root of the tree. To be efficient, this procedure requires that the data associated with the nodes in a generation occupy a consecutive region in the address-space. This eliminates the need for synchronization between the worker threads, because each worker thread can deduce its own portion based on its thread-id and the address-range of the generation. To orchestrate the worker teams, the master thread only needs to keep account of the address-ranges. Technically, this is accomplished by iterating over a vector of offsets (algorithm 5.2).

Algorithm 5.2 : Range-based parallel pruning

Input : $\mathcal{T}$, $\mathbf{t}$, $\mathbf{z}$, $\boldsymbol{\Theta}$
Output : $S_M(\mathbf{t}, \mathbf{z}, \boldsymbol{\Theta})$
Data :
`/* A pre-calculated vector with starting offsets for each generation: */`
1 $Range = \left[0, N, N + |G_1|, N + |G_1| + |G_2|, ..., M - 1, M\right]_{K+1}$;
`/* a vector of` M `elements */`
2 $State \longleftarrow [0, ..., 0]_M$;
`/* The master thread iterates over the generations: */`
3 **foreach** $k \in \{1, ..., K\}$ **do**
  `/* The master thread starts a team of worker threads running equal portions of the following loop: */`
4 **foreach** $j \in \{Range[k] + 1, ..., Range[k + 1]\}$ **do**
5 $State[j] \longleftarrow R_j\left(\left\{State[i] : i \in Desc(j)\right\}, \mathbf{t}, \mathbf{z}, \boldsymbol{\Theta}\right)$;
6 **return** $State[M]$;

In algorithm 5.2, the number of synchronization points is reduced to the number of generations, K. In balanced trees, K would increase logarithmically with N and, for big N, the tree would be split into a few generations of many nodes (fig. 5.1a). Conversely, in strongly unbalanced trees, K would tend to increase linearly with N and the tree would be split into many generations of a few nodes (fig. 5.1b). This would result in low parallel speed-up and excessive synchronization cost for both, the queue-based and the range-based algorithms. Also noteworthy is the fact that algorithm 5.2 reduces the number of synchronization points at the cost of some parallelization. If each worker thread gets assigned to an approximately equal number of nodes in a generation and if a few of the nodes take much longer time to process than the rest, then most of the worker threads would have to wait until the last node in the generation has been processed.

These and other subtleties (Supplementary Information) indicate that there is no "one size fits all" strategy when it comes to maximizing parallel speed-up. The framework provides two ways to deal with these: (a) allowing the user to choose a parallelization mode before executing a pruning operation on a given tree and data; (b) providing a mode "auto", in which the framework compares the execution time of different pruning algorithms during

the first several calls on a given tree and data, choosing the fastest one for all subsequent calls.

5.2.3 *The SPLITT library*

We provide SPLITT in the form of an open source C++ library licensed under version 3.0 of the GNU Lesser General Public License (LGPL v3.0) and available on `https://github.com/venelin/SPLITT.git`. In its current implementation, the library uses the C++11 language standard, the standard template library (STL) and the OpenMP standard for parallel processing. The library is designed as a set of C++ template classes, generic with respect to the application specific details, such as the types of input data, model parameters and definitions of the node states, S_i, and visit-node functions, R_i. The library defines two layers (fig. S1):

- a framework layer defining the main logical and data structures. These include a linear algorithm for initial reordering and splitting of the input tree into generations of nodes, which can be visited in parallel, both during post-order as well as pre-order traversal, and a growing collection of pre-order and post-order traversal algorithms, targeting different parallelization modes (e.g. queue-based versus range-based parallelization) on different computing devices (currently implemented for CPUs only).
- a user layer at which the user of the library must write a `CustomTraversalSpecification` class defining all typedefs and methods of the interface `TraversalSpecification`. The methods that should be defined by the user are:
  - `SetParameter(par)`: sets parameter values, such as model parameters, prior to tree-traversal.
  - `InitNode(i)`: called for each node, i, at the beginning of the traversal; performs node-specific initialization, based on the parameter-values and the input data; can be executed both, sequentially or in parallel, depending on the selected parallelization mode; this function is the perfect place to define the calculation of node-specific state fields or other node-specific data, which depend on the parameters, the tree and the input data but do not depend on the state/data associated with other nodes;
  - `VisitNode(i)`: called for the root (in pre-order traversals only) and for every internal and tip node, i, (both, pre-order and post-order traversals) after `InitNode(i)` and either after `VisitNode(j)` and `PruneNode(j, i)` has been called for each $j \in Desc(i)$ in post-order traversals, or after `VisitNode(`$Parent(i)$`)` has been called in pre-order traversals. This method is suitable for implementing the logic in the the function R_i, depending on the parameters, the input data, and the state of the nodes, on which i's state depends.
  - `PruneNode(i, i_parent)`: called solely in post-order traversals for every node, i, after the call to `VisitNode(i)` and before calling `VisitNode(`$Parent(i)$`)`. This method is suitable for updating fields associated with $Parent(i)$ before it gets visited. It is logically equivalent to leave the implementation of `PruneNode(i, i_parent)` empty and have the implementation of `VisitNode(i)` consult the states of its daughter nodes.
  - `StateAtRoot()`: returns the state associated with the root of the tree.

The bridge between the two layers is provided by an object of the `TraversalTask` template class (fig. S1). Once the `TraversalSpecification` implementation has been written, the

user instantiates a `TraversalTask` object passing the tree and the input data as arguments. This triggers the creation of the internal objects of the framework, i.e. an `OrderedTree` object maintaining the order in which the nodes are processed and a `PreOrderTraversal` or a `PostOrderTraversal` object implementing different parallelization modes of the two traversal types. In the ideal use-case, the `TraversalTask`'s `TraverseTree()` method will be called repeatedly, varying the model parameters, the input data and branch lengths on a fixed tree topology. This encompasses all scenarios where a model is fitted to a fixed tree and data, e.g. ML or Bayesian PCM inference.

5.3 RESULTS

We evaluated the performance of the SPLITT library using a univariate and a multivariate Phylogenetic Ornstein-Uhlenbeck Mixed Model (POUMM) as a showcase. Previously, we and other authors have used this model as an estimator of pathogen trait heritability in large HIV cohorts (Bertels et al., 2017; Mitov and Stadler, 2018). A detailed description of the POUMM can be found in Supplementary Information and in (Mitov and Stadler, 2018). The univariate POUMM was implemented in the R-package `POUMM`, based on a quadratic polynomial representation of the log-likelihood (Supplementary Information). The multivariate POUMM version was implemented in a new R-package, `PCMBaseCpp`, using a multivariate generalization of the quadratic polynomial representation described elsewhere (manuscript in preparation). The two packages perform post-order tree traversal by linking to SPLITT through the package `Rcpp` (Eddelbuettel and Sanderson, 2014). The POUMM is a suitable model for a comparative benchmark, because a number of R-packages provide similar OU-based phylogenetic models, using C++ for the likelihood implementation. These include, among others, `geiger` (Pennell et al., 2014) and `diversitree` (FitzJohn, 2012) for the univariate case and `Rphylopars` (Goolsby, Bruggeman, and Ané, 2016) for the multivariate case.

We used the R-package `apTreeshape` (Bortolussi et al., 2012) to generate tree topologies of sizes $N \in \{100; 1000; 10,000; 100,000\}$. To generate the trees, we used the function `rtreeshape()` with a `biased` model. A parameter p in this model controls the disproportion of branching rates for the left and right lineages starting from a given parent node. For each N, we used four settings for p as follows:

1. $p = 0.5$ corresponding to equal left and right branching rates and resulting in balanced trees;
2. $p = 0.1$ corresponding to unbalanced trees in which one of any two sibling branches (sharing the same parent node) splits at rate $p = 0.1$, while the other splits at rate $p' = 1 - p = 0.9$ (time units are arbitrary, so we can assume that the rates correspond to splitting probabilities per unit time).
3. $p = 0.01$ corresponding to very unbalanced trees (splitting rates of $p = 0.01$ and $p' = 0.99$ for any couple of sibling branches;
4. $p = 0.01/N$ corresponding to a ladder-like tree (see fig. 5.1b, 5.2).

This resulted in a total of 16 topologies (trees for $N = 1,000$ shown on fig. 5.2). For each topology, random branch lengths were assigned overwriting the default branch lengths of 1 assigned by `rtreeshape()`. Since the OU-implementations in the current `diversitree` and `Rphylopars` versions do not support non-ultrametric trees, each tree was ultrametrized (adjusting branch lengths so that all tips have the same root-tip distance). For each tree, we generated random trait-values using random parameters of the POUMM model.

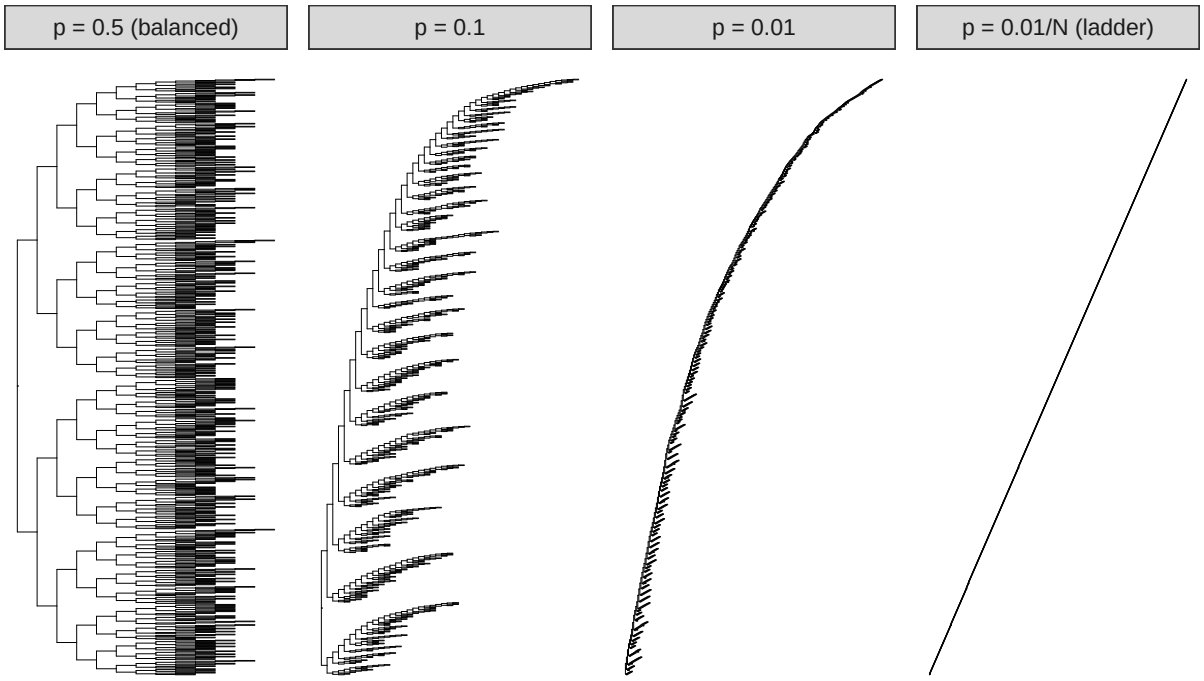


Figure 5.2: Test tree topologies for $N = 1,000$. For visualization purpose, all branch lengths have been set to 1, whereas the random branch lengths were used in the benchmarks.

5.3.1 *Time for preprocessing the tree*

Each of the tested packages implements a preprocessing step initializing cached data-structures that are re-used during likelihood calculation. In the case of `SPLITT`, this is the constructor of the class `TraversalTask` (fig. S1); in the case of `diversitree`, this is the function `make.ou`; in the case of `geiger`, this is the internal function `bm.lik`. We note that the time for creating the cache structure is not important in scenarios of fitting Gaussian phylogenetic models to a fixed tree and data (created once, at the beginning of the inference process). However, these times become important in the case when the tree topology is inferred together with the model parameters from trait and sequence alignment data.

We measured the preprocessing time on the 16 trees (table 5.1). The times scaled linearly with the size of the tree for the packages using the SPLITT library (`POUMM` and `PCMBaseCpp`) and for `diversitree`. For these packages the time was not affected by the unbalancedness of the tree. For `geiger`, we observed longer times, both for bigger N as well as for more unbalanced trees. For $N = 100,000$ and $p = 0.01/N$, both, `diversitree` and `geiger` failed with a `stack-overflow` error. The relatively short times for the SPLITT-based `POUMM` and `PCMBaseCpp` packages indicate that `SPLITT` could potentially be used for phylogenetic inference.

5.3.2 *Time for POUMM likelihood calculation*

To measure the likelihood calculation time, we ran performance benchmarks on a MacBook Pro laptop (Retina, 15-inch, Late 2013) 4 CPU cores and on the "Euler" scientific cluster (https://scicomp.ethz.ch/wiki/Euler) with up to 24 CPU cores. Here, we comment on the calculation times on MacBook Pro, noting that the times on Euler for up to 4 CPU cores were nearly equal (Supplementary Information, figs S3-S7).

We distinguish the different implementations according to the following criteria:

Table 5.1: Times for tree-preprocessing in milliseconds.

N	Implementation	p=0.5	p=0.1	p=0.01	p=0.01/N
100	geiger	5	6	9	9
100	diversitree	4	4	4	4
100	SPLITT	2	2	2	1
1,000	geiger	18	26	78	414
1,000	diversitree	20	20	22	30
1,000	SPLITT	3	2	3	3
10,000	geiger	358	449	1,345	355,396
10,000	diversitree	207	211	227	1,338
10,000	SPLITT	14	13	13	15
100,000	geiger	20,215	21,629	36,349	-
100,000	diversitree	2,421	2,619	2,883	-
100,000	SPLITT	130	131	131	140

- Number of traits: we distinguish between univariate implementations, i.e. `geiger`, `diversitree` and `POUMM`, and multivariate implementations, i.e. `Rphylopars` and `PCMBaseCpp`. For the multivariate implementations, we measured the time for 1, 4, 8 and 16 traits.
- Mode: denotes whether the implementation is single threaded using one physical core of the CPU - serial, or multi-threaded, running as many threads as there are physical CPU cores - parallel;
- Order: denotes the order in which the prune-able nodes are processed. We tested three possible orders: postorder (only for Mode=serial) - the nodes are processed sequentially; queue-based (only for Mode=parallel) - the nodes are processed in parallel as they enter the queue (see algorithm 5.1), synchronized thread access to the queue; range-based (only for Mode=parallel) - the nodes in each pruning generation are processed in order of their allocation in memory, no need for a synchronized access to a queue (see algorithm 5.2).
- Implementation: the R-package and the back-end used (R or C++).

The resulting times for the univariate implementations running on the MacBook Pro computer are shown on fig. 5.3.

On small trees of 100 tips, the fastest univariate POUMM implementations were the serial C++ implementations from the packages `POUMM` and `diversitree` (about 0.03 ms); the range-based parallel implementation was nearly as fast on balanced trees ($p = 0.5$) but was progressively slower on unbalanced trees. The geiger implementation was nearly an order of magnitude slower (0.2 ms). The POUMM queue-based parallel implementation was nearly 100 times slower (nearly 2 ms), presumably due to the excessive synchronization overhead. The serial R implementation from the diversitree package was the slowest (above 2 ms), which was expected, since the R interpreter is notorious for its slow speed compared to compiled languages like C++. On bigger balanced trees ($N > 100$, $p = 0.5$), the range-based parallel implementation took over, reaching up to $4\times$ speed-up with respect to the range-based serial implementation, up to $5\times$ speed-up with respect to the postorder serial implementation and

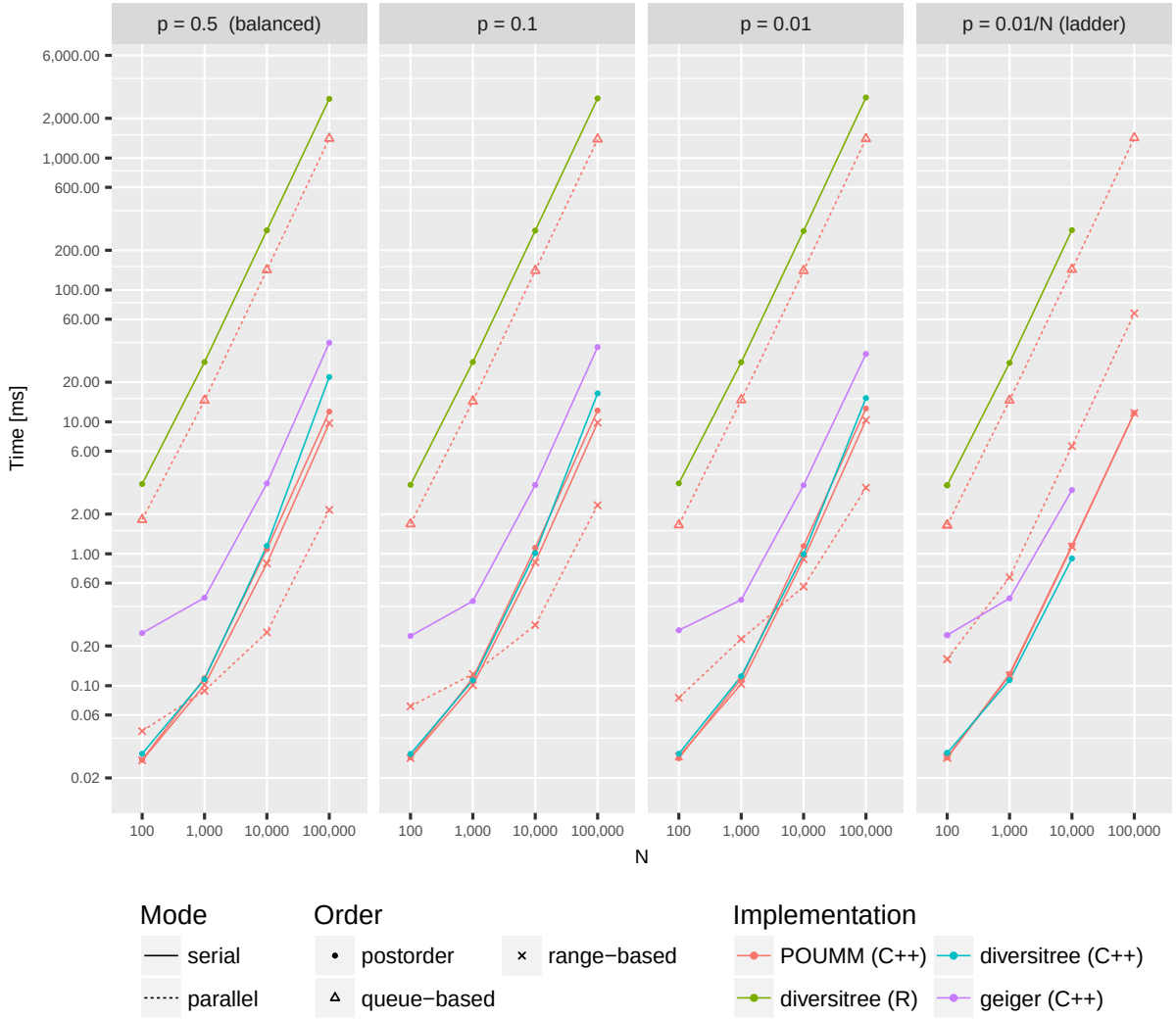


Figure 5.3: Likelihood calculation times for univariate R and C++ implementations of the POUMM model on a MacBook Pro late 2013 computer (processor Intel(R) Core(TM) i7-4850HQ CPU @ 2.30GHz with four physical cores). Both, the $x-$axis denoting the number of tips in the tree and the $y-$axis denoting the calculation time in milliseconds are on a log-10 scale. Panels from left to right correspond to different tree topologies with left-most panel corresponding to a perfectly balanced tree and right-most panel corresponding to a ladder tree, see also fig. 5.2.

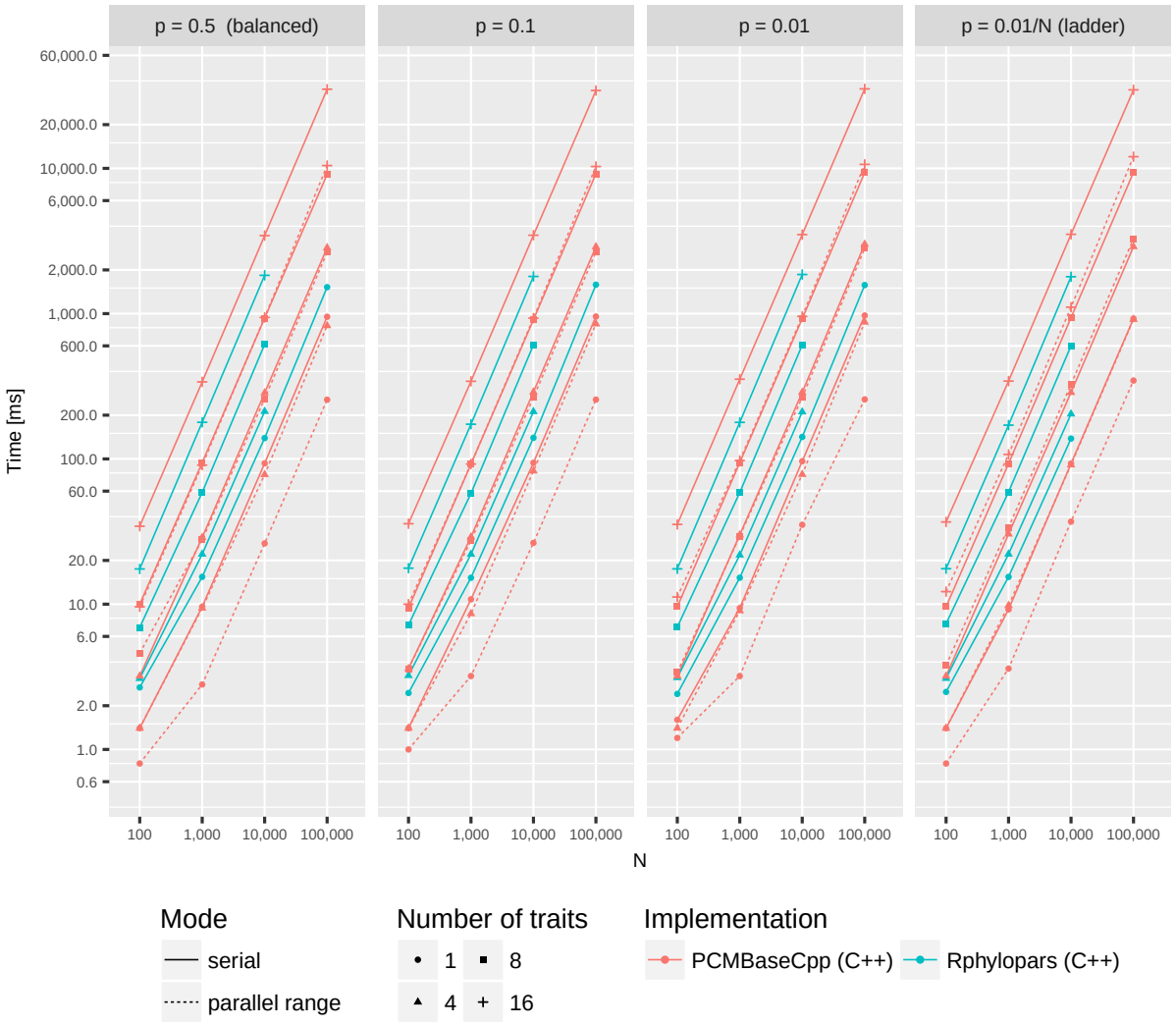


Figure 5.4: Likelihood calculation times for multivariate C++ implementations of the POUMM model on a MacBook Pro late 2013 computer (processor Intel(R) Core(TM) i7-4850HQ CPU @ 2.30GHz with four physical cores). The panel layout and the $x, y-$axes are the same as on fig. 5.3. For simplicity, only serial and parallel range modes are shown, noting that the parallel queue mode had slightly slower times compared to the parallel range mode.

up to 10× speed-up with respect to the `diversitree` serial C++ implementation. This reveals a consistent speed-up for all trees except the ladder tree, where parallelization of the internal nodes is not possible (see fig. 5.1b and 5.2). The time for the other serial implementations and the POUMM queue-based parallel implementation scaled up linearly with N.

The times for the multivariate implementations running on the MacBook Pro computer are shown on fig. 5.4. For these implementations, the likelihood calculation times were about two orders of magnitude higher compared to the univariate implementations. This is due to slow algebraic operations, for example arithmetic division in the univariate case as opposed to matrix inversion in the multivariate case.

5.3.3 *Parallel speedup*

The parallel speed-ups for the Euler cluster benchmark for univariate implementations and for multivariate implementations with 16 traits are shown on figs. S8 and 5.6 (see also figs. S8-S10, for multivariate implementations with 1, 4 and 8 traits).

For univariate implementations, the parallel speed-up is negligible for trees of less than 1000 tips aND for highly unbalanced tree (fig. S8. The parallel speed-up becomes noticeable for large balanced trees, peaking at 10x for a balanced tree of 100,000 tips, running on 20 CPU cores (fig. S8). The above behaviour is explained by the fact that the init-node and visit-node operations in the univariate case are very fast relative to the thread-management operations. Also noteworthy is the fact that even on balanced trees above 100,000 tips, the parallel efficiency, i.e. the ratio of the parallel speed-up and the number of parallel cores, drops below 50% when running on more than 20 CPU cores. This suggests a possible competition between the CPU cores for a limited resource such as the processor cache or the memory bandwidth.

For the multivariate implementations, the init-node and visit-node operations are computationally more intensive. This is why we observe substantial parallel speed-up on the smallest as well as the most unbalanced trees trees (fig. 5.6). However, for all multivariate cases, we observe a decline in parallel speed-up with more than 12 CPU cores (fig. 5.6). The most reasonable explanation for this is competition between the CPU cores for a limited hardware resource.

5.4 DISCUSSION

The examples in this article focused on Gaussian models of continuous trait evolution (Supplementary Information), yet, SPLITT can in principle be used for any algorithm that runs a pre-order or post-order tree traversal. For example, another family of models where SPLITT could be used are models of structured populations. When calculating the likelihood for a phylogenetic tree under a structured birth-death model, the calculations proceed in a pruning fashion (Kühnert et al., 2016) and may be improved with respect to speed using our approach. However, the structured coalescent likelihood for a tree is a function of all coexisting lineages even for approximate methods (Müller, Rasmussen, and Stadler, 2017), and thus a pruning formulation is not available.

We did not develop examples of pre-order traversal. One such example is the simulation of traits evolving along the tree, which can be used for validation and approximate inference of phylogenetic models. In complex phylogenetic comparative models, where an exact calculation of the likelihood is elusive or computationally intractable, it is possible to use simulations of trait evolution along the tree for approximate likelihood calculation (Kutsukake and Innan, 2013) or approximate Bayesian computation (ABC) (Slater et al., 2012). Both approaches are computationally intensive and could benefit from parallel execution using SPLITT.

We should not omit mentioning other software libraries implementing parallel likelihood computation of different Markov models of sequence evolution. For example, several high level tools for ML and Bayesian tree inference, e.g. Bouckaert et al. (2014), Drummond et al. (2012), and Ronquist and Huelsenbeck (2003), use the library BEAGLE which distributes the computation for the independent sites of the sequence alignment among multiple CPU or GPU cores (Ayres et al., 2012). SPLITT operates on a different level, namely, it parallelizes the computation for independent lineages in the tree. Both approaches are interesting because they fit well to different sizes of the input data - while BEAGLE achieves significant parallel speed-ups in long alignments comprising many thousands nucleotide or codon columns

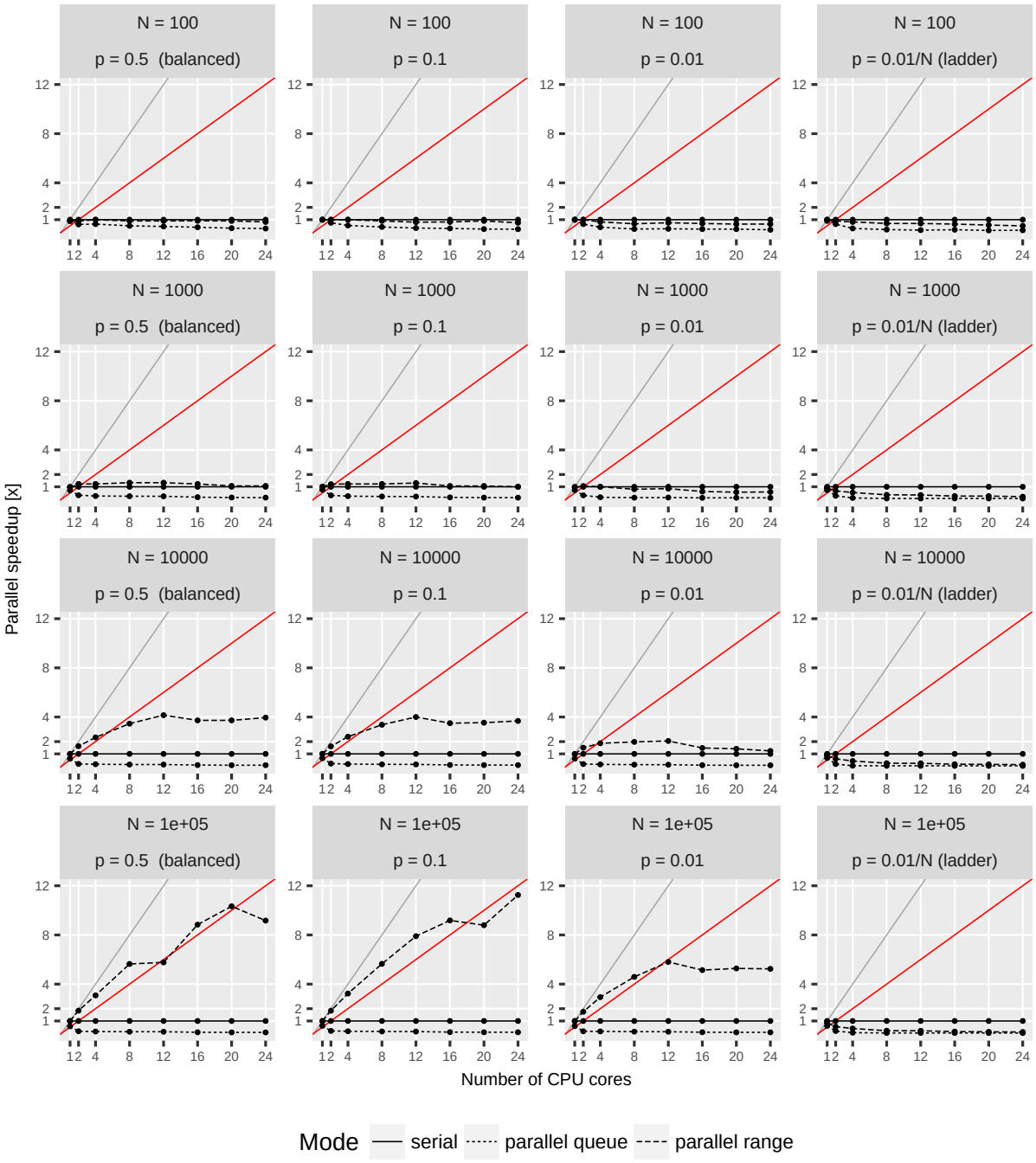


Figure 5.5: Parallel speed-up for the univariate POUMM implementation on the Euler cluster (package `POUMM`). The grey and red lines denote, the expected speed-up at 100% and 50% parallel efficiency, respectively. Horizontally, the panels correspond to the different tree topologies, see also fig. 5.2. Vertically, the panels correspond to the different tree-sizes.

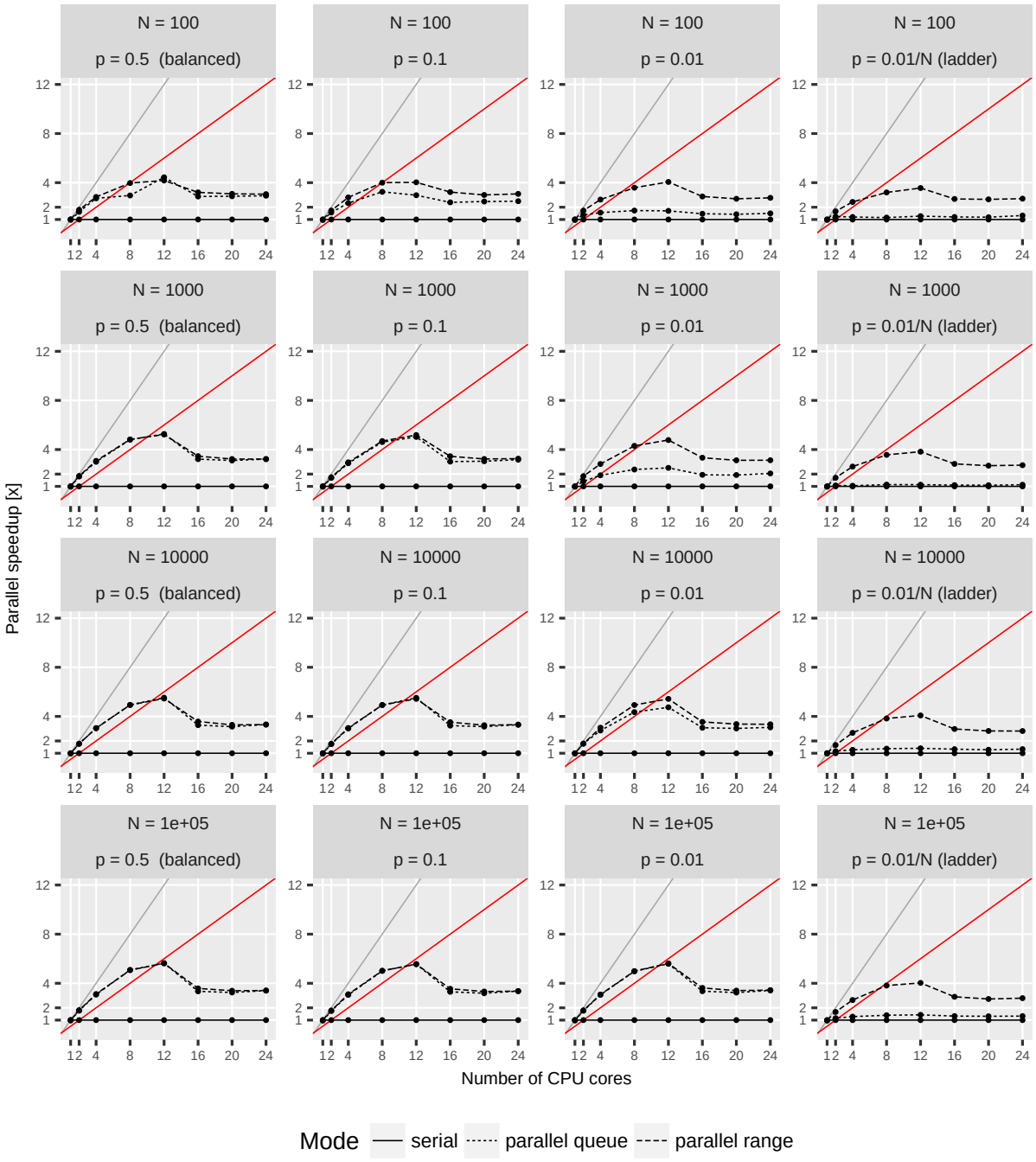


Figure 5.6: Parallel speed-up for the multivariate POUMM implementation (package `PCMBaseCpp`) on the Euler cluster. The grey and red lines denote, the expected speed-up at 100% and 50% parallel efficiency, respectively. Horizontally, the panels correspond to the different tree topologies, see also fig. 5.2. Vertically, the panels correspond to the different tree-sizes.

(Ayres et al., 2012), SPLITT is better suited to shorter alignments of potentially many thousands of species.

Based on the performance benchmarks, we conclude that with the current implementation of SPLITT, running on the above-mentioned hardware, the parallel speed-up from parallel tree traversal is up to one order of magnitude using up to 20 CPU cores. For comparison, Ayres et al., 2012 report up to 33x speed-up with BEAGLE on a codon alignment of less than 20 sequences with 6000 codons, using double floating point precision on a GPU of 600 cores. A future GPU-based extension of SPLITT would show if it can reach higher levels of parallel speed-up and efficiency. Reaching higher speed-up of the Bayesian inference, though, is possible if the parallel traversal likelihood calculation is combined with a general purpose adaptive Metropolis sample. An example application of this combined approach to real data is given in Supplementary Information.

5.4.1 *Outlook*

The past decade has seen a rapid advance in the production of multi-core processors. At the same time, it appears that the maximum clock frequency of a single processing unit is approaching the maximum achievable for semi-conductor based architectures. In parallel with this development on the hardware side, the volume of sequence data and the size of phylogenetic trees is growing exponentially. For instance, in less than five years the size of phylogenetic trees used for calculating the heritability of HIV virulence has increased from a few hundreds to several thousand patients (Alizon et al., 2010; Bachmann et al., 2017; Bertels et al., 2017; Blanquart et al., 2017; Hodcroft et al., 2014). This motivates the development of novel parallel algorithms capitalizing on the multi-core technology. The parallel tree traversal library, SPLITT, enables parallel computation for a vast set of phylogenetic models, facing the challenges of increasing model complexity and volumes of data in phylogenetic analysis.

5.5 SUPPLEMENTARY MATERIAL

Data from the performance benchmarks and simulations for technical correctness is available on the SPLITT github page `https://github.com/venelin/SPLITT`. The `POUMM` package and user guide is available at `https://github.com/venelin/POUMM`. The `PCMBaseCpp` package is available at `https://github.com/venelin/POUMM`.

5.6 FUNDING

V.M. and T.S. thank ETH Zürich for funding. T.S. is supported in part by the European Research Council under the 7th Framework Programme of the European Commission (PhyPD: Grant Agreement Number 335529).

5.7 ACKNOWLEDGEMENTS

We thank Dr. Krzysztof Bartoszek for valuable insights on the Ornstein-Uhlenbeck process.

APPENDIX

5.A EXAMPLES OF USING THE PARALLEL TREE TRAVERSAL FRAMEWORK

In this section, we show example usages of the parallel traversal framework. In each of these examples, we solve a particular problem, such as calculating the likelihood of a continuous Markov model for a categorical or a continuous trait. In terms of the framework, the task boils down to formulating the node states $S_j(\mathbf{t}, \mathbf{z}, \boldsymbol{\Theta})$ and the recursive functions R_j satisfying rules (1) and (2).

5.A.1 *Example 1. Gaussian models of continuous trait evolution*

Ho and Ané (2014a) noticed that the computational complexity in multivariate Gaussian and some non-Gaussian models concentrates in the calculation of determinants $|\mathbf{V}_{\boldsymbol{\Theta}}|$ and quadratic quantities of the form $\mathbf{Q}_{\boldsymbol{\Theta}} = \mathbf{X}'_{\boldsymbol{\Theta}}\mathbf{V}_{\boldsymbol{\Theta}}^{-1}\mathbf{Y}_{\boldsymbol{\Theta}}$, where $\mathbf{V}_{\boldsymbol{\Theta}}$ represents the variance covariance matrix expected under the model specified by $\boldsymbol{\Theta}$ and the matrices $\mathbf{X}_{\boldsymbol{\Theta}}$ and $\mathbf{Y}_{\boldsymbol{\Theta}}$ represent centered observed data at the tips in the tree. For example, in the case of Brownian motion and Ornestein-Uhlenbeck models, the log-likelihood function is equal to the log-density of a multivariate Gaussian distribution:

$$\ln f(\mathbf{z}|\boldsymbol{\Theta}) = -\frac{1}{2}\left(N\ln(2\pi) + \ln|\mathbf{V}_{\boldsymbol{\Theta}}| + (\mathbf{z} - \bar{\mathbf{z}}_{\boldsymbol{\Theta}})'\mathbf{V}_{\boldsymbol{\Theta}}^{-1}(\mathbf{z} - \bar{\mathbf{z}}_{\boldsymbol{\Theta}})\right), \tag{S1}$$

where $\mathbf{V}_{\boldsymbol{\Theta}} = \boldsymbol{\Sigma}$ and $\bar{\mathbf{z}}_{\boldsymbol{\Theta}} = \bar{\mathbf{z}}$ (table S1).

Ho and Ané (2014a) developed a pruning algorithm which allows to calculate $|\mathbf{V}_{\boldsymbol{\Theta}}|$ and $\mathbf{Q}_{\boldsymbol{\Theta}}$ simultaneously and without constructing or allocating the matrix $\mathbf{V}_{\boldsymbol{\Theta}}$ in memory, provided $\mathbf{V}_{\boldsymbol{\Theta}}$ has a "3-point structure". Then, they showed several examples of Gaussian models such as Brownian motion and Ornstein-Uhlenbeck, as well as non-Gaussian models, such as phylogenetic logistic and Poisson regression, where $\mathbf{V}_{\boldsymbol{\Theta}}$ is or can be "converted" to a 3-point structured matrix (discussed later). Adapting the notation from (Ho and Ané, 2014a,

Table S1: Population properties at the tips of the phylogeny under BM and OU models and their mixed counterparts The acronyms are: PBM - Phylogenetic Brownian motion (without non-heritable component); PMM - Phylogenetic Mixed Model (adding a non-heritable component to PBM); POU - Phylogenetic Ornstein-Uhlenbeck (without non-heritable component), also known as "Hansen's model" or Single Stationary Peak (SSP); POUMM - Phylogenetic Ornstein-Uhlenbeck Mixed Model (adding a non-heritable component to the POU model. Expressions for the OU-models were adapted from (Hansen, 1997). $\mu_{\mathbf{\Theta},i}$: expected value at tip i; $\mathbf{V}_{\mathbf{\Theta},ii}$: expected variance for tip i; $\mathbf{V}_{\mathbf{\Theta},ij}$: expected covariance of the values of tips i and j.

	PBM	PMM	POU	POUMM
$\mathbf{\Theta}$:	$< g_M, \sigma >$	$< g_M, \sigma, \sigma_e >$	$< g_M, \alpha, \theta, \sigma >$	$< g_M, \alpha, \theta, \sigma, \sigma_e >$
$\mu_{\mathbf{\Theta},i}$:	g_M	g_M	$e^{-\alpha h_i} g_M + \left(1 - e^{-\alpha h_i}\right) \theta$	$e^{-\alpha h_i} g_M + \left(1 - e^{-\alpha h_i}\right) \theta$
$\mathbf{V}_{\mathbf{\Theta},ii}$:	$\sigma^2 h_i$	$\sigma^2 h_i + \sigma_e^2$	$\frac{\sigma^2}{2\alpha} \left(1 - e^{-2\alpha h_i}\right)$	$\frac{\sigma^2}{2\alpha} \left(1 - e^{-2\alpha h_i}\right) + \sigma_e^2$
$\mathbf{V}_{\mathbf{\Theta},ij}$:	$\sigma^2 h_{(ij)}$	$\sigma^2 h_{(ij)}$	$\frac{\sigma^2}{2\alpha} e^{-\alpha d_{ij}} \left(1 - e^{-2\alpha h_{(ij)}}\right)$	$\frac{\sigma^2}{2\alpha} e^{-\alpha d_{ij}} \left(1 - e^{-2\alpha h_{(ij)}}\right)$

p. 399), we define the node states as $S_j(\mathbf{t}, \mathbf{z}, \mathbf{\Theta}) = \left\langle p_{A,j}, p_j, \hat{\mu}_{Y,j}, \tilde{\mu}'_{X,j}, \ln|\mathbf{V}|_j, \mathbf{Q}_j \right\rangle$. The recursive functions R_j follow immediately from points 1 and 2 of the algorithm (Ho and Ané, 2014a):

$$\begin{cases} S_j(\mathbf{t}, \mathbf{z}, \mathbf{\Theta}) = \langle \; p_{A,j} = 0, \\ \qquad\qquad\quad p_j = \frac{1}{t_j}, \\ \qquad\qquad\quad \hat{\mu}_{Y,j} = \mathbf{y}_{\mathbf{\Theta},j}, \\ \qquad\qquad\quad \tilde{\mu}'_{X,j} = \mathbf{x}'_{\mathbf{\Theta},j}, \\ \qquad\qquad\quad \ln|\mathbf{V}|_j = \ln t_j, \\ \qquad\qquad\quad \mathbf{Q}_j = \mathbf{x}'_{\mathbf{\Theta},j} \mathbf{y}_{\mathbf{\Theta},j} \quad \rangle & \text{if } j \leq N \\ S_j(\mathbf{t}, \mathbf{z}, \mathbf{\Theta}) = \langle \; p_{A,j} = \sum_{i \in Desc(j)} p_i, \\ \qquad\qquad\quad p_j = \frac{p_{A,j}}{1 + t_j p_{A,j}}, \\ \qquad\qquad\quad \hat{\mu}_{Y,j} = \sum_{i \in Desc(j)} \frac{p_i}{p_A} \hat{\mu}_{Y,i}, \\ \qquad\qquad\quad \tilde{\mu}'_{X,j} = \sum_{i \in Desc(j)} \frac{p_i}{p_A} \tilde{\mu}'_{X,i}, \\ \qquad\qquad\quad \ln|\mathbf{V}|_j = \sum_{i \in Desc(j)} \ln|\mathbf{V}|_i + \ln(1 + t_j p_{A,j}), \\ \qquad\qquad\quad \mathbf{Q}_j = \sum_{i \in Desc(j)} \mathbf{Q}_i + \ln(1 + t_j p_{A,j}) \quad \rangle & \text{otherwise.} \end{cases} \tag{S2}$$

The caveat in applying the 3-point algorithm is that except for BM models, the matrix $\mathbf{V}_{\mathbf{\Theta}}$ does not necessarily satisfy the 3-point condition (Ho and Ané, 2014a). As the authors show, it is still possible to use the algorithm in that case, provided that $\mathbf{V}_{\mathbf{\Theta}}$ satisfies a "generalized 3-point condition" (Ho and Ané, 2014a). More precisely, in most of their examples, the authors showed that there exist a transformation of the branch lengths, $\tilde{\mathbf{t}}$, diagonal matrices $\mathbf{D}_1$ and $\mathbf{D}_2$ with non-zero diagonal elements and a 3-point structured matrix $\tilde{\mathbf{V}}_{\mathbf{\Theta}}$, such that $\tilde{\mathbf{V}}_{\mathbf{\Theta}}$ is equal to the variance-covariance on the tree $\tilde{\mathcal{T}}$ with the transformed branch lengths and

$\mathbf{V_\Theta} = \mathbf{D}_1\tilde{\mathbf{V}}_\Theta\mathbf{D}_2$. If so, the algorithm is applied to $\tilde{\mathbf{V}}_\Theta$ using $\tilde{\mathbf{t}}$ and transformed data $\tilde{\mathbf{X}} = \mathbf{D}_2^{-1}\mathbf{X}$, $\tilde{\mathbf{Y}} = \mathbf{D}_1^{-1}\mathbf{Y}$. Then the quadratic form of interest, $\mathbf{Q}_\Theta$, would be equal to the resulting quadratic form at the root, $\mathbf{Q}_M$ and the determinant $|\mathbf{V}_\Theta|$ is obtained by the formula:

$$|\mathbf{V}_\Theta| = |\mathbf{D}_1||\tilde{\mathbf{V}}_\Theta||\mathbf{D}_2| \tag{S3}$$

5.A.2 *Example 2: The phylogenetic Ornstein-Uhlenbeck mixed model*

Here, we describe a phylogenetic Ornstein-Uhlenbeck mixed model of continuous trait evolution, which we and other authors have used previously to analyze the evolution of set-point viral load in HIV patients (Bachmann et al., 2017; Bertels et al., 2017; Blanquart et al., 2017; Mitov and Stadler, 2016, 2018).

5.A.2.1 *The model*

Consider a continuous trait evolving independently along the lineages of a phylogenetic tree, $\mathcal{T}$ with branch lengths $\mathbf{t}$. The phylogenetic Ornstein-Uhlenbeck mixed model (POUMM) decomposes the trait value as a sum of a non-heritable component, e, and a genetic component, g, which (i) evolves continuously according to an Ornstein-Uhlenbeck (OU) process along branches; (ii) gets inherited by the branches descending from each internal node. In biological terms, g is a genotypic value (Lynch and Walsh, 1998) that evolves according to random drift with stabilizing selection towards a global optimum; e is a non-heritable component, which can be interpreted in different ways, depending on the application, i.e. a measurement error, an environmental contribution, a residual with respect to a model prediction, or the sum of all these. The OU-process acting on g is parameterized by an initial genotypic value at the root, g_M, a global optimum, θ, a selection strength, $\alpha>0$, and a random drift unit-time standard deviation, σ. Denoting by W_t the standard Wiener process (Grimmett and Stirzaker, 2001), the evolution of the trait-value, $z(t)$, along a given lineage of the tree is described by the equations:

$$z(t) = g(t) + e \tag{S4}$$
$$dg(t) = \alpha[\theta - g(t)]dt + \sigma dW_t \tag{S5}$$
$$g(0) = g_M, \tag{S6}$$

The stochastic differential equation S5 defines the OU-process, which represents a random walk tending towards the global optimum θ with stronger attraction for bigger difference between $g(t)$ and θ (Ornstein and Zernike, 1919; Uhlenbeck and Ornstein, 1930). The model assumptions for e are that they are independent and identically distributed (i.i.d.) normal with mean 0 and standard deviation σ_e at the tips. Any process along the tree that gives rise to this distribution at the tips may be assumed for e. For example, in the case of epidemics, a newly infected individual is assigned a new e-value which represents the contribution from its immune system and this value can change or remain constant throughout the course of infection. In particular, the non-heritable component e does not influence the behavior of the OU-process $g(t)$. Thus, if we were to simulate trait values z on the tips of a phylogenetic tree $\mathcal{T}$, we could first simulate the OU-process from the root to the tips to obtain g, and then add the white noise e (i.e. an i.i.d. draw from a normal distribution) to each simulated g value at the tips. The POUMM represents an extension of the phylogenetic mixed model

(PMM) (Housworth, Martins, and Lynch, 2004; Lynch, 1991), since, in the limit $\alpha \to 0$, the OU-process converges to a Brownian motion (BM) with unit-time standard deviation σ. Both, the POUMM and the PMM, define an expected multivariate normal distribution for the trait values at the tips. The mean vectors and the variance-covariance matrices of these distributions are written in table S1. Note that the trait expectation and variance for a tip i depends on its height (h_i), and the trait covariance for a pair of tips (ij) depends on the height of their mrca ($h_{(ij)}$), and, in the case of POUMM, on their patristic distance (d_{ij}) (table S1).

5.A.2.2 *Calculating the POUMM likelihood*

Here we describe two ways to calculate the POUMM likelihood using a post-order traversal of the tree, which can be easily incorporated with the framework. The first approach is based on the generalized 3-point structure algorithm (Ho and Ané, 2014a). This approach has the caveat that it requires a model-specific branch-length transformation. The second approach is based on direct integration over the ancestor genotypic values at the internal nodes, capitalizing on a recurrent quadratic polynomial formula. Previously, a similar integration technique has been described in (FitzJohn, 2012). The advantage of the quadratic polynomial representation described here is that it can be generalized to multivariate OU models as well as a more general class of Gaussian models (in preparation).

GENERALIZED 3-POINT STRUCTURE OF THE POUMM VARIANCE-COVARIANCE MATRIX

The POUMM likelihood is defined as the multivariate probability density of an observed vector $\mathbf{z}$ at the tips of $\mathcal{T}$ for given model parameters $\boldsymbol{\Theta} =< g_M, \alpha, \theta, \sigma, \sigma_e >$:

$$\ell\ell(\boldsymbol{\Theta}) = \ln(f(\mathbf{z}|\mathcal{T}, \mathbf{t}, \boldsymbol{\Theta})). \tag{S7}$$

The probability density function, f is multivariate Gaussian with mean vector $\bar{}_{\boldsymbol{\Theta}}$ and variance-covariance matrix $\mathbf{V}_{\boldsymbol{\Theta}}$ written in table S1. Since $\mathbf{V}_{\boldsymbol{\Theta}}$ has a generalized 3-point structure (Ho and Ané, 2014a), we can apply the recursion in eq. S2, upon a transformation of the branch lengths and the data. This is obtained through adapting the transformation for an non-mixed OU-model in a ultrametric tree (Ho and Ané, 2014a) to accommodate the non-heritable variance:

$$\tilde{t}_i = \frac{\sigma^2}{2\alpha}\left[e^{2\alpha T}\left(e^{2\alpha h_i} - e^{2\alpha h_{Parent(i)}}\right)\right] + \frac{\sigma_e^2}{e^{2\alpha u_i}}\delta(i \leq N) \quad \text{for } i \in \{1, ..., M-1\} \tag{S8}$$

$$\tilde{\mathbf{X}}_i = \tilde{\mathbf{Y}}_i = \frac{z_i - \mu_i}{e^{\alpha u_i}} \quad \text{for } i \in \{1, ..., N\}, \tag{S9}$$

where $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{Y}}$ are identical N-vectors, T is the maximum tip-height in the tree and $u_i = T - h_i$ for $i \in \{1, ..., N\}$. After running the post-order traversal, using eq. S2 as a visit-node operation, we apply eq. S3, to obtain $|\mathbf{V}_{\boldsymbol{\Theta}}|$ and eq. S1 to obtain the log-likelihood.

We note that the branch transformation (eq. S8) can be done "locally" on every branch, using pre-calculated heights of the parent and daughter nodes connected by the branch. Thus, it is safe to include the transformation in the visit-node operation and the parallelization of pruning would not suffer. Otherwise, the transformation would have had to be done in a preprocessing step. Again, this is a model specific consideration.

A QUADRATIC POLYNOMIAL REPRESENTATION OF THE POUMM LOG-LIKELIHOOD We begin by defining for each nodes j states, $S_j(\mathbf{t}, \mathbf{z}, \mathbf{\Theta})$, and recursive functions R_j that allow the calculation of the likelihood, $\ell\ell(\Theta)$, from S_M. It turns out that $\ell\ell(\mathbf{\Theta})$ has a simple representation as a quadratic polynomial of g_M (root state), which can be obtained by pruning–wise integration over the unobserved internal node states, g_i, progressing from the tips to the root. We formalize this idea in the following theorem:

Theorem 1 (Recurrent quadratic polynomial representation of the POUMM log-likelihood)**.** *For $\alpha \geq 0$, a real θ and non-negative σ and σ_e, the POUMM log-likelihood can be expressed as a quadratic polynomial of g_M:*

$$\ell\ell(\mathbf{\Theta}) = a_M g_M^2 + b_M g_M + c_M, \tag{S10}$$

where $a_M < 0$, b_M and c_M are real coefficients. We denote by $u(\alpha, t)$ the function:

$$u(\alpha, t) := \begin{cases} \alpha/(1 - e^{\alpha t}), & \text{for } \alpha > 0 \\ -1/t, & \text{for } \alpha = 0 \end{cases} \tag{S11}$$

Then, the coefficients in eq. S10 can be expressed with the following recurrence relation:
1. For $j \in \{1, ..., N\}$ (tips):

$$a_j = -\frac{1}{2\sigma_e^2};\ b_j = \frac{z_j}{\sigma_e^2};\ c_j = -\frac{z_j^2}{2\sigma_e^2} - \ln\sqrt{2\pi\sigma_e^2} \tag{S12}$$

2. For $j > N$ (internal nodes) or $j = M$ (root):

$$\begin{aligned} a_j &= \sum_{i \in Desc(j)} \frac{a_i u(\alpha, 2t_i)}{u(\alpha, 2t_i) - \alpha + \sigma^2 a_i} \\ b_j &= \sum_{i \in Desc(j)} \frac{u(\alpha, 2t_i)\left[2\theta a_i (e^{\alpha t_i} - 1) + b_i e^{\alpha t_i}\right]}{u(\alpha, 2t_i) - \alpha + \sigma^2 a_i} \\ c_j &= \sum_{i \in Desc(j)} \Bigg\{ c_i + \alpha t_i - \frac{0.25\, b_i^2 \sigma^2}{-\alpha + a_i\sigma^2 + u(\alpha, 2t_i)} - \\ &\qquad 0.5 \ln\left(\frac{-\alpha + a_i\sigma^2 + u(\alpha, 2t_i)}{u(\alpha, 2t_i)}\right) + \\ &\qquad \frac{\alpha\theta\left[a_i\theta - (b_i + a_i\theta)e^{\alpha t_i}\right]}{u(\alpha, t_i) + (-\alpha + a_i\sigma^2)\left(1 + e^{\alpha t_i}\right)} \Bigg\}. \end{aligned} \tag{S13}$$

Proof. Induction from the tips to the root of the tree.

- *Basis:* For a tip-node i, $\mathcal{T}_i$ is the trivial tree consisting of this tip-node only and the pdf of $\mathbf{z}_i$, conditioned on the unobservable genotypic value g_i, is given by the normal pdf with mean g_i and variance σ_e^2. This pdf can be written as:

$$\begin{aligned} f(\mathbf{z}_i | g_i; \sigma_e) &= \mathcal{N}(z_i; g_i, \sigma_e^2) \\ &= \frac{1}{\sqrt{2\pi\sigma_e^2}} e^{-\frac{(z_i - g_i)^2}{2\sigma_e^2}} \\ &= e^{-\frac{1}{2\sigma_e^2} g_i^2 + \frac{z_i}{\sigma_e^2} g_i - \frac{z_i^2}{2\sigma_e^2} - 0.5\ln(2\pi\sigma_e^2)} \end{aligned} \tag{S14}$$

By defining $a_i = -\frac{1}{2\sigma_e^2}$, $b_i = \frac{z_i}{\sigma_e^2}$ and $c_i = -\frac{z_i^2}{2\sigma_e^2} - 0.5\ln(2\pi\sigma_e^2)$ and taking the natural logarithm of the pdf we obtain the log-likelihood representation from eq. S10, where $a_M < 0$, b_M and c_M can be calculated from the observed value z_i and the model parameter σ_e.

- *Inductive hypothesis:* Assume that for an internal node j, the statement of the theorem has been proven for all subtrees $\mathcal{T}_i$, $i \in Desc(j)$.
- *Inductive step:* Assuming that g_j is known, we consider the OU process starting from g_j and parametrized by α and σ. Under this process, the expected distribution at time t_i is normal with mean $\mu_{ji} = e^{-\alpha t_i} g_j + (1 - e^{-\alpha t_i})\theta$ and variance $\sigma_{ji}^2 = (1 - e^{-2\alpha t_i})\frac{\sigma^2}{2\alpha}$. Then, the probability of $\mathbf{z}_i$ given g_j is given by the integral

$$\begin{aligned}
f(\mathbf{z}_i|\boldsymbol{\Theta}, t_i, g_j) &= \int_{-\infty}^{\infty} f(g_i|\boldsymbol{\Theta}, t_i, g_j) \times e^{a_i g_i^2 + b_i g_i + c_i} dg_i \\
&= \int_{-\infty}^{\infty} \mathcal{N}\left[g_i; \mu_{ji}, \sigma_{ji}^2\right] \times e^{a_i g_i^2 + b_i g_i + c_i} dg_i \\
&= \int_{-\infty}^{\infty} e^{(p_{ji}+a_i) g_i{}^2 + (q_{ji}+b_i) g_i + (r_{ji}+c_i)} dg_i \ , \ \text{where} \\
p_{ji} &= -\frac{1}{2\sigma_{ji}^2} = -\frac{\alpha e^{2\alpha t_i}}{\sigma^2\left(e^{2\alpha t_i}-1\right)} \\
q_{ji} &= \frac{\mu_{ji}}{\sigma_{ji}^2} = \frac{2\alpha e^{\alpha t_i}\left[g_j + \theta\left(e^{\alpha t_i}-1\right)\right]}{\sigma^2\left(e^{2\alpha t_i}-1\right)} \\
r_{ji} &= -\frac{\mu_{ji}^2}{2\sigma_{ji}^2} - \frac{1}{2}\ln(2\pi\sigma_{ji}^2) \\
&= -\frac{\alpha\left[g_j + \theta\left(e^{\alpha t_i}-1\right)\right]^2}{\sigma^2\left(e^{2\alpha t_i}-1\right)} - \frac{1}{2}\ln\left(\frac{\pi\sigma^2\left(1-e^{-2\alpha t_i}\right)}{\alpha}\right)
\end{aligned} \tag{S15}$$

We notice that p_{ji}, q_{ji} and r_{ji} in eq. S15 are not defined in the case of BM ($\alpha = 0$). In this case, we take the limit for $\alpha \to 0$ represented by the case $\alpha = 0$ of function $u(\alpha, t)$ (eq. S11). By substituting $u(\alpha,\ t)$ in the expressions for p_{ji}, q_{ji}, r_{ji} (eq. S15) we obtain:

$$\begin{aligned}
p_{ji} &= \frac{e^{2\alpha t_i} u(\alpha, 2t_i)}{\sigma^2} \\
q_{ji} &= -\frac{u(\alpha, 2\,t_i)\left[g_j + \theta\left(e^{\alpha t_i}-1\right)\right]}{\sigma^2} \\
r_{ji} &= \frac{u(\alpha, 2\,t_i)\left[g_j + \theta\left(e^{\alpha t_i}-1\right)\right]^2}{\sigma^2} - \frac{1}{2}\ln\left(-\frac{\pi\sigma^2}{u(\alpha, 2t_i)e^{2\alpha t_i}}\right).
\end{aligned} \tag{S16}$$

Since $a_i < 0$ and, for positive t and $\alpha \in [0, \infty)$, $u(\alpha,\ t)$ accepts strictly negative values in the interval $[-1/t, 0)$, the integral in eq. S15 has a closed form solution:

$$\begin{aligned}
&\int_{-\infty}^{\infty} e^{(p_{ji}+a_i)g_i^2+(q_{ji}+b_i)g_i+(r_{ji}+c_i)} dg_i \\
&= \exp\left[\frac{-(q_{ji}+b_i)^2}{4(p_{ji}+a_i)} + (r_{ji}+c_i) + \ln\left(\sqrt{\frac{\pi}{-(p_{ji}+a_i)}}\right)\right] \\
&= e^{a_{ji}g_j^2+b_{ji}g_j+c_{ji}}\ , \text{ where} \\
a_{ji} &= \frac{a_i u(\alpha,2t_i)}{u(\alpha,2t_i)-\alpha+\sigma^2 a_i} \\
b_{ji} &= \frac{u(\alpha,2t_i)(e^{\alpha t_i}(2\theta a_i+b_i)-2\theta a_i)}{u(\alpha,2t_i)-\alpha+\sigma^2 a_i} \\
c_{ji} &= c_i + \alpha t_i - \frac{0.25\, b_i^2\sigma^2}{-\alpha+a_i\sigma^2+u(\alpha,2t_i)} - \\
&\qquad 0.5\ln\left(\frac{-\alpha+a_i\sigma^2+u(\alpha,2t_i)}{u(\alpha,2t_i)}\right) + \frac{\alpha\theta\left[a_i\theta-(b_i+a_i\theta)e^{\alpha t_i}\right]}{u(\alpha,t_i)+(-\alpha+a_i\sigma^2)\left(1+e^{\alpha t_i}\right)}
\end{aligned} \tag{S17}$$

In eq. S17 above, $a_{ji} < 0$ because it is a fraction with a positive nominator and a negative denominator (note that $a_i < 0$ by the inductive hypothesis and $u(\alpha, 2t_i) < 0$ by definition). Since the vectors $\mathbf{z}_i$, $i \in Desc(j)$, are conditionally independent given g_j, the conditional pdf of $\mathbf{z}_j$ factorizes as:

$$\begin{aligned}
f(\mathbf{z}_j|\boldsymbol{\Theta}, g_j, \mathcal{T}_j) &= \prod_{i\in Desc(j)} f(\mathbf{z}_i|\boldsymbol{\Theta}, t_i, g_j) \\
&= \prod_{i\in Desc(j)} e^{a_{ji}g_j^2+b_{ji}g_j+c_{ji}} \\
&= \exp\left[\left(\sum_{i\in Desc(j)} a_{ji}\right) g_j^2 + \left(\sum_{i\in Desc(j)} b_{ji}\right) g_j + \sum_{i\in Desc(j)} c_{ji}\right].
\end{aligned} \tag{S18}$$

By denoting $a_j = \sum_{i\in Desc(j)} a_{ji}$, $b_j = \sum_{i\in Desc(j)} b_{ji}$ and $c_j = \sum_{i\in Desc(j)} c_{ji}$ and noticing that $a_j < 0$ as a sum of negative terms, we have proven the inductive step and, thus, the theorem.

□

5.A.3 *Example 3: Models of categorical trait evolution*

Before moving on to the main show-case of this work, it is worthy mentioning that SPLITT can be readily applied to any pruning-wise calculation, including calculating the likelihoods of categorical trait models. Consider a trait taking values in $\{0, 1\}$ evolving independently along the lineages of a phylogenetic tree, $\mathcal{T}$ with branch lengths $\mathbf{t}$. A continuous-time Markov model can be used to characterize the transitions of the trait value along each branch (Felsenstein, 1983; Pagel, 1994). This model assumes constant rates of change from 0 to 1, q_{01} and

from 1 to 0, q_{10}, representing the probability that the change has occurred during an infinitesimal interval of time. These rates are used to define a rate matrix:

$$\mathbf{Q} = \begin{bmatrix} -q_{01} & q_{01} \\ q_{10} & -q_{10} \end{bmatrix}. \tag{S19}$$

Given $\mathbf{Q}$, the transition probability matrix $\mathbf{P}(t)$ for an arbitrary long period t is given by

$$\mathbf{P}(t) = \begin{bmatrix} P_{00}(t) & P_{01}(t) \\ P_{10}(t) & P_{11}(t) \end{bmatrix} = \mathbf{C} \begin{bmatrix} e^{\lambda_1 t} & 0 \\ 0 & e^{\lambda_2 t} \end{bmatrix} \mathbf{C}^{-1} \tag{S20}$$

where λ_i are the eigenvalues of $\mathbf{Q}$ and $\mathbf{C}$ is a matrix, which's i^{th} column represents the i^{th} eigenvector of $\mathbf{Q}$ (Pagel, 1994). Assuming that the value at the root is known to be z_M, we want to find the probability with which the model specified by the parameters $\mathbf{\Theta} = (q_{01}, q_{10})$ generates an N-vector of values, $\mathbf{z}$ observed at the tips. This represents the conditional likelihood $L_{\mathcal{T}}(\mathbf{t}, \mathbf{z}, \mathbf{\Theta}, z_M)$. The pruning algorithm for calclulating L relies on calculating the "fragmentary" likelihood $L_i(b) = P(\mathbf{z}_i | z_i = b; \mathbf{\Theta})$ for each node i and each $b \in \{0, 1\}$ (Felsenstein, 1983). In terms of the framework, we define the state $S_j(\mathbf{t}, \mathbf{z}, \mathbf{\Theta})$ of a node j as the pair $< L_j(0), L_j(1) >$. Following eq. 4 in (Felsenstein, 1983), the recursive R_j are given by:

$$S_j(\mathbf{t}, \mathbf{z}, \mathbf{\Theta}) = \begin{cases} \left\langle \delta(z_j = 0), \delta(z_j = 1) \right\rangle & \text{if } j \text{ is a tip} \\ \left\langle \prod_{i \in Desc(j)} \left[\sum_{z_i} P_{0 z_i}(t_i) L_i(z_i) \right], \prod_{i \in Desc(j)} \left[\sum_{z_i} P_{1 z_i}(t_i) L_i(z_i) \right] \right\rangle & \text{if } j \text{ is internal,} \end{cases}$$

(S21)

where we use the Kronecker delta function $\delta(x = y)$ equalling to 1 if $x = y$ and 0, otherwise. In the above eqation S21, the values $L_i(z_i)$ are available from the descendants' states S_i. Finally, the conditional likelihood $L_{\mathcal{T}}(\mathbf{t}, \mathbf{z}, \mathbf{\Theta}, z_M)$ is given by $L_M(z_M)$, which is one of the two members in S_M.

The above model can be extended to a multivariate case, such as calculating the probability of a nucleotide or aminoacid sequence alignment as is the case in (Felsenstein, 1983). Suppose that there are p nucleotide sites, which are evolving independently. Then, the state for a node j would represent a $p \times 4$ matrix

$$S_j(\mathbf{t}, \mathbf{z}, \mathbf{\Theta}) = \begin{bmatrix} L_j^{(1)}(A) & L_j^{(1)}(C) & L_j^{(1)}(T) & L_j^{(1)}(G) \\ \vdots & \vdots & \vdots & \vdots \\ L_j^{(p)}(A) & L_j^{(p)}(C) & L_j^{(p)}(T) & L_j^{(p)}(G) \end{bmatrix}, \tag{S22}$$

where the letters A, C, T and G denote the nucleotides and the superscript in parentheses denotes a site in the alignment. To define the recursive functions R_j, equation S21 can be extended to accomodate one row of S_j (four possible values instead of two) and evaluated p times to obtain the full state.

The model can also be extended to support correlated evolution between the sites. As shown in (Pagel, 1994), this involves extending the rate matrix **Q** to embed transition rates between pairs, triplets or higher order combinations of sites in the sequence. Accounting for correlated evolution between combinations of sites dramatically increases the computational complexity, but does not present a conceptual change from the point of view of the pruning operation. Thus, accommodating such models in the framework, although involved technically, should not present a conceptual challenge.

5.B OTHER PARALLELIZATION STRATEGIES

5.B.1 *Hybrid parallel/sequential strategies*

An important problem occurring with all parallel pruning strategies is that the number of lineages tends to decreases exponentially towards the root of the tree. As a result, if the original thread-team consisted of numerous threads each one reserving one of multiple processing cores, there will be many idle threads/cores as the pruning approaches the root. While this issue could potentially be solved at the level of the multi-threading back-end, it is possible to implement a hybrid pruning strategy similar to the rake-compression algorithm described in (Reif, 1989). Reif (1989) introduce two operations: rake (could be seen as the parallel calculation on the nodes in one generation) and compress (compression of chains). For example, on a ladder tree (fig. 5.1b), after visiting the generation of tips, one obtains a chain: the chain should be processed sequentially on one thread (compressed), thus, reducing the synchronization thread-starvation issues.

5.C DESIGN OF THE SPLITT LIBRARY

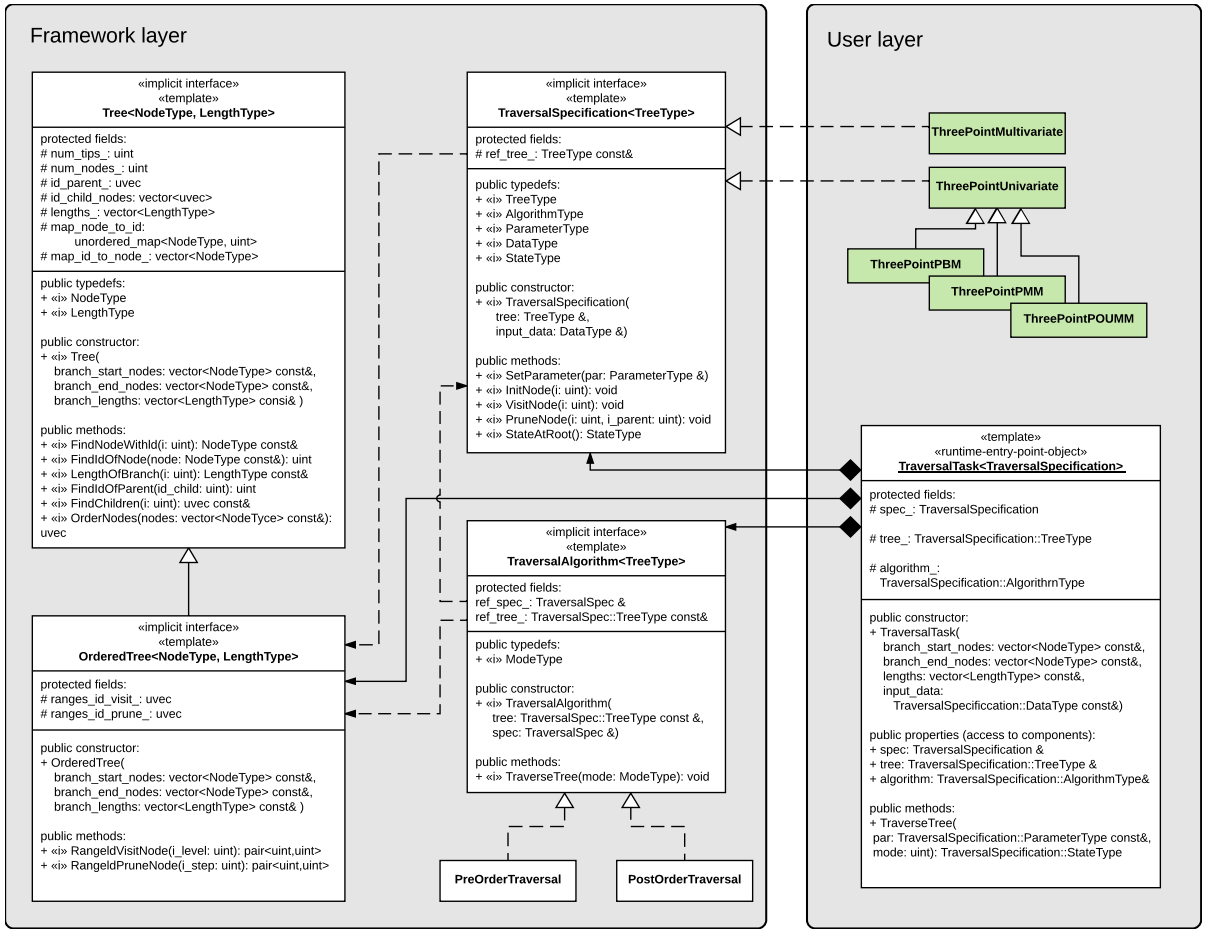


Figure S1: **A class diagram of the SPLITT library** In the framework layer, the class `TraversalSpecification` defines the application-specific data types and logic; the class `Tree` serves as a base-class implementing common tree operations, such as constructing a tree from a list of branches, checking the validity of the input (e.g. lack of cycles or isolated branches), finding the parent and the descendants of a node, etc; the class `OrderedTree` maintains the order of the nodes in a tree so that they can be split in contiguous generations for parallel post-order or pre-order traversal; the class `TraversalAlgorithm` serves as a base class and an implicit interface for its two subclasses implementing the two supported types of tree traversal: `PreOrderTraversal` and `PostOrderTraversal`. At the user layer are the user defined implementations of the `TraversalSpecification` interface (shown in green) and an instance of the `TraversalTask` class, which constructs all necessary internal objects and serves as a runtime entry point to the framework.

5.D THE POUMM R-PACKAGE

We implement the POUMM model in the form of an R-package called `POUMM`, which embeds the SPLITT library as an Rcpp module. Before model fitting, the user can choose from different POUMM parametrizations and prior settings (function `specifyPOUMM`). A set of standard generic functions, such as `plot`, `summary`, `logLik`, `coef`, etc., provide means to assess the quality of a fit (i.e. MCMC convergence, consistence between ML and MCMC fits) as well as various inferred properties, such as high posterior density (HPD) intervals (more details in the package user guide).

5.D.1 *Model inference*

We implement maximum likelihood and Bayesian inference of the POUMM parameters, $\boldsymbol{\Theta}$, using the L-BFGS-R convex optimization algorithm (R-function `optim`) and a variant of the Random Walk Metropolis (RWM) Markov Chain Monte Carlo (MCMC) sampling (Metropolis et al., 1953). This combined inference capitalizes on two practical ideas:

- A MCMC has higher chance to converge to the target posterior distribution faster if it has been started from a previously estimated MLE;
- If an MCMC encounters a point in the parameter space that has higher likelihood than a previously inferred MLE, running maximum likelihood optimization from that point is more likely to find the global likelihood optimum.

An important step in RWM is the choice of a proposal (jump) distribution shape matrix used as a scaling factor on each next proposal in the Metropolis algorithm. Choosing the shape matrix with respect to the scale and the correlation structure of the parameter space minimizes the number of iterations needed for MCMC convergence and mixing. Thus, numerous variants of the RWM have been proposed, performing "on-the-fly" adaptation of the shape matrix based on what has been "learned" about the parameter space from the past RWM iterations (Haario, Saksman, and Tamminen, 2001; Vihola, 2012). Of these variants, we chose the adaptive Metropolis sampling with coerced acceptance rate, because it is shown to be robust with respect to the posterior distribution, it performs a relatively cheap adaptation of the shape (Vihola, 2012) and it has an implementation in the R within the package `adaptMCMC` (Scheidegger, 2012).

The fitting of the POUMM model was implemented as a pipeline including the following steps:

1. Perform three MLE searches using the R-function `optim` and the L-BFGS-B method (Byrd et al., 1995), starting from three randomly chosen points in parameter space;
2. Run three MCMC chains as follows: (i) a chain sampling from the prior distribution; (ii) a chain sampling from the posterior distribution and started from the MLE found in step 1; (iii) a chain sampling from the posterior distribution and started from a random point in parameters space.
3. If the parameter tuple of highest likelihood sampled in the MCMC has a likelihood higher than the MLE found in step 1, repeat the MLE search starting from that parameter tuple;

By running MLE first and starting an MCMC chain from the MLE candidate, we increase the chance that at least one of the MCMCs would converge faster to the posterior distribution. By comparing the posterior samples from two MCMCs initiated from different starting

points, it can be assessed whether the MCMCs have converged to the true posterior. We do this quantitatively by the use of the Gelman-Rubin convergence diagnostic (Brooks and Gelman, 1998) implemented in the R-package coda (Plummer et al., 2006). Values of the Gelman-Rubin (G.R.) statistic significantly different from 1 indicate that at least one of the two posterior samples deviates significantly from the true posterior distribution. By visual comparison of posterior density with prior desnity plots, it is possible to assess whether the data contains information differring from the prior knowledge for a given parameter. In step 3, we capitalize on the chance that the MCMCs have explored a wider region of the parameter space than the likelihood optimization.

5.D.2 *Technical correctness*

To validate the correctness of the Bayesian POUMM implmentation, we used the method of posterior quantiles (Cook, Gelman, and Rubin, 2006). In this method, the idea is to generate samples from the posterior quantile distributions of selected model parameters (or functions thereof) by means of numerous "replications" of simulation followed by Bayesian parameter inference. In each replication, "true" values of the model parameters are drawn from a fixed prior distribution and trait-data is simulated under the model specified by these parameter values. We perform these simulations on a fixed tree of size $N = 4000$. Then, the to-be-tested software is used to produce a posterior distribution of parameters based on the simulated trait-data. Next, the posterior quantiles of the "true" parameter values (or functions thereof) are calculated from the corresponding posterior samples generated by the to-be-tested software. By running multiple independent replications on a fixed prior, it is possible to generate large samples from the posterior quantile distributions of the individual model parameters, as well as any derived quantities. Assuming correctness of the simulations, any statistically significant deviation from uniformity of these posterior quantile samples indicates an error in the to-be-tested software (Cook, Gelman, and Rubin, 2006).

Two phylogenetic trees were used for the simulations:

- Ultrametric (BD, $N = 4000$) - an ultrametric birth-death tree of 4000 tips generated using the TreeSim R-package (Boskova, Bonhoeffer, and Stadler, 2014; Stadler et al., 2013) (function call: `sim.bd.taxa(4000, lambda = 2, mu = 1, frac = 1, complete = FALSE)`);
- Non-ultrametric (BD, $N = 4000$) - a non-ultrametric birth-death tree of 4000 tips generated using the TreeSim R-package (Boskova, Bonhoeffer, and Stadler, 2014; Stadler et al., 2013) (function call: `sim.bdsky.stt(4000, lambdasky = 2, deathsky = 1, timesky=0)`).

Simulation scenarios of 2000 replications were run using the prior distribution $< g_M, \alpha, \theta, \sigma, \sigma_e > \sim \mathcal{N}(5, 25) \times \text{Exp}(0.1) \times \mathcal{U}(2, 8) \times \text{Exp}(0.4) \times \text{Exp}(1)$. The goal of using this prior was to explore a large enough subspace of the POUMM parameter space, while keeping MCMC convergence and mixing within reasonable time (runtime up to 30 minutes for two MCMCs of 10^6 adaptive Metropolis iterations at target acceptance rate of 1%). From the above prior, we drew a sample of $n = 2000$ parameter tuples, $\{\boldsymbol{\Theta}^{(1)}, ..., \boldsymbol{\Theta}^{(n)}\}$, which were used as replication seeds. For a given $\boldsymbol{\Theta}^{(i)}$, we simulate genotypic values $\mathbf{g}^{(i)}(\mathcal{T}, \boldsymbol{\Theta}^{(i)})$ according to an OU-branching process with initial value $g_M^{(i)}$ and parameters $\alpha^{(i)}$, $\theta^{(i)}$, $\sigma^{(i)}$. Then, we add random white noise ($\sim \mathcal{N}(0, \sigma_e^{2(i)})$) to the genotypic values at the tips, to obtain the final trait values $\mathbf{z}^{(i)}$.

For the two simulated trees, we executed a total of $2 \times 2000 = 4000$ replications. The resulting posterior quantile distributions for the each tree are shown on Fig. S2. We notice that the posterior quantiles for all relevant parameters are uniformly distributed. This is

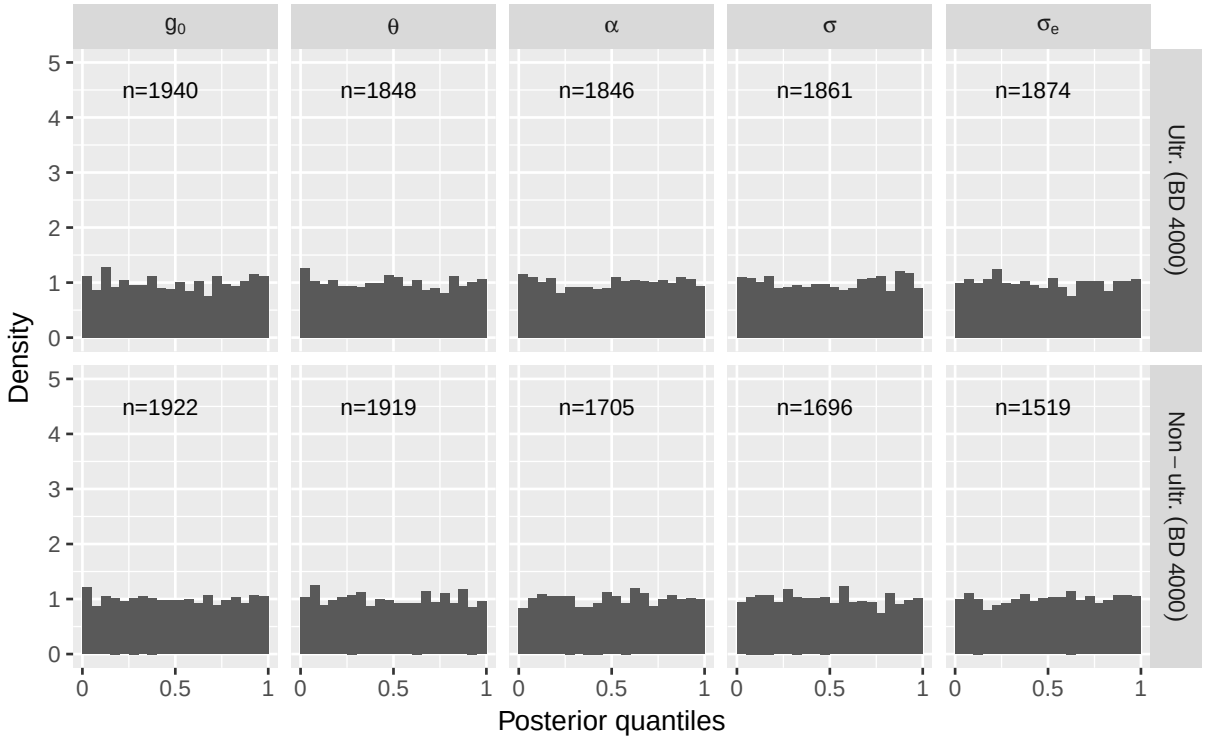


Figure S2: Posterior quantiles from simulations on a ultrametric and a non-ultrametric tree ($N = 4000$). The number n at the top of each histogram denotes the number of replications out of 2000 which reached acceptable MCMC convergence and mixing after one million iterations. Uniformity was confirmed using a Kolmogorov-Smirnov test which was insignificant for all parameters (P-value above 0.1).

confirmed visually by the corresponding histograms (fig. S2), as well as statistically, by a non-significant p-value from a Kolmogorov-Smirnov uniformity test at the 0.01 level. This observation validates the technical correctness of the software.

5.E SUPPLEMENTARY RESULTS FROM THE PERFORMANCE BENCHMARKS

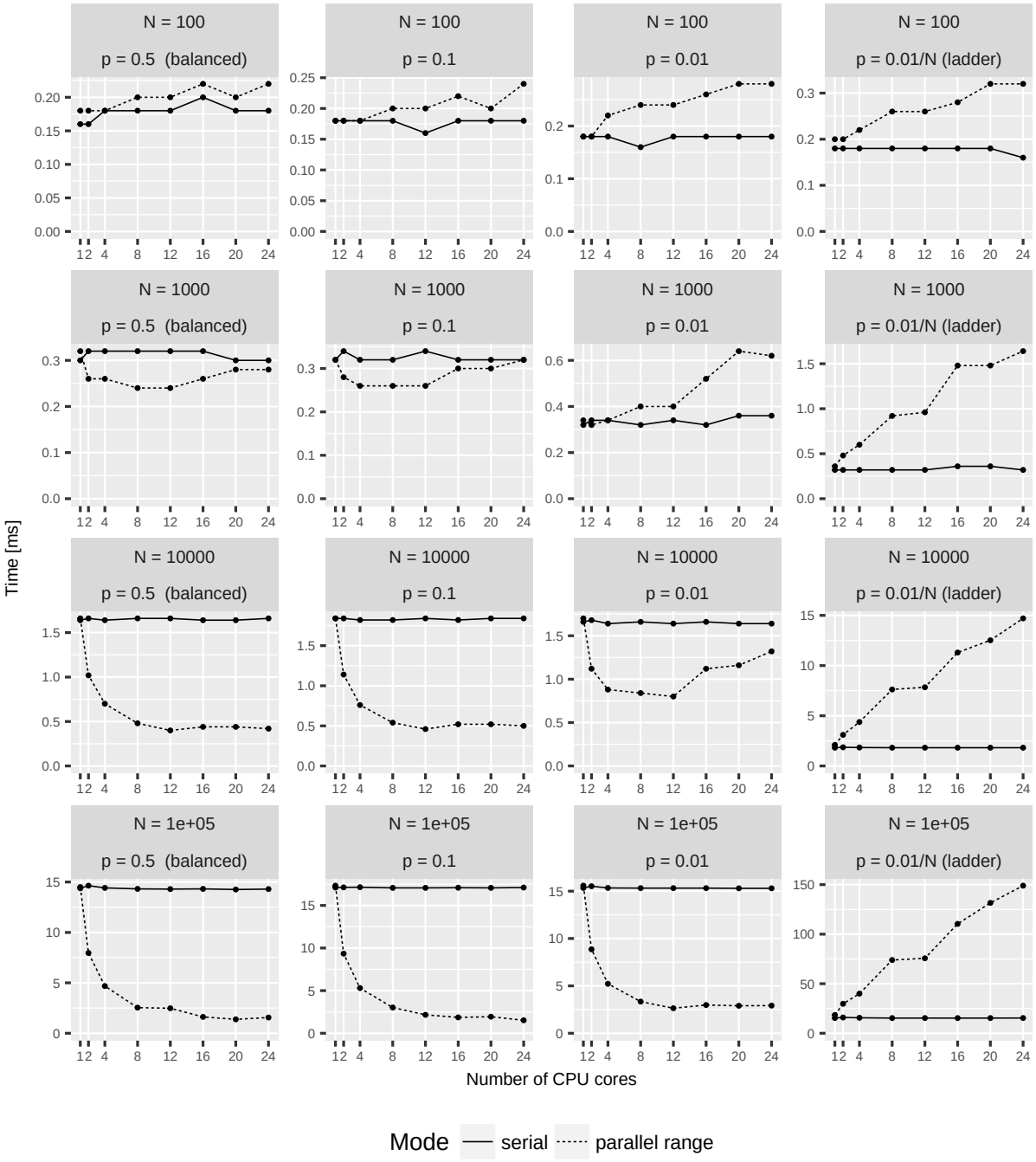


Figure S3: Likelihood calculation times for the univariate POUMM implementation (package `POUMM`) on Euler cluster (a single shared memory node with Intel(R) Xeon(R) CPU E5-2680 v3 @ 2.50GHz running 24 physical cores). Both, the $x-$axis denoting the number of cores, and the $y-$axis denoting the calculation time in milliseconds are on the linear scale. Horizontally, the panels correspond to the different tree topologies, see also fig. 5.2. Vertically, the panels correspond to the different tree-sizes. For visualization purpose, only the times for the serial postorder and the fastest parallel algorithm (parallel range-based) are shown. The times for the parallel queue-based implementation were significantly higher.

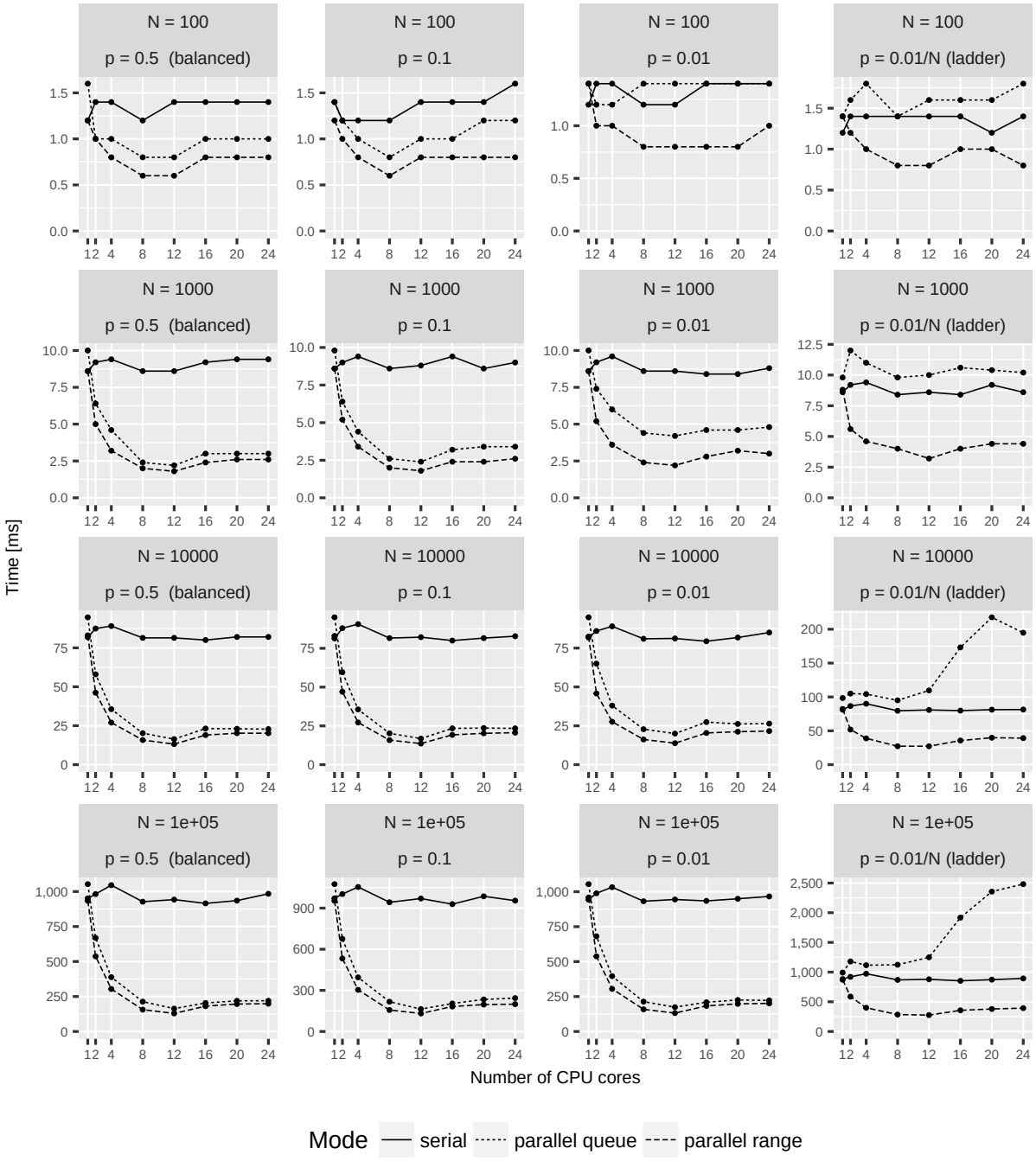


Figure S4: Likelihood calculation times for the multivariate POUMM implementation (package `PCMBaseCpp`) with 1 trait on Euler cluster (a single shared memory node with Intel(R) Xeon(R) CPU E5-2680 v3 @ 2.50GHz running 24 physical cores). Both, the $x-$axis denoting the number of cores, and the $y-$axis denoting the calculation time in milliseconds are on the linear scale. Horizontally, the panels correspond to the different tree topologies, see also fig. 5.2. Vertically, the panels correspond to the different tree-sizes.

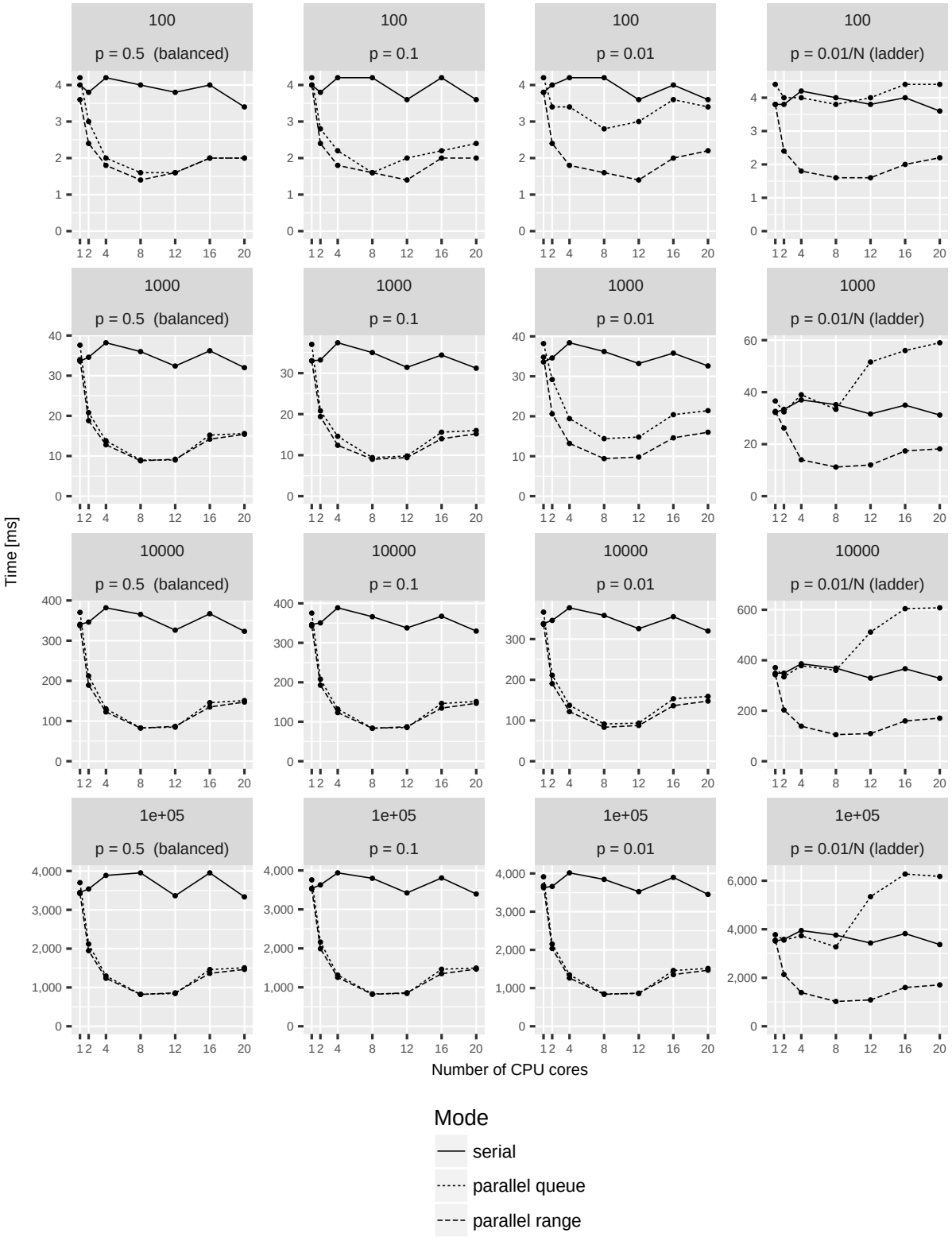


Figure S5: Likelihood calculation times for the multivariate POUMM implementation (`PCMBaseCpp`) with 4 traits on Euler cluster (a single shared memory node with Intel(R) Xeon(R) CPU E5-2680 v3 @ 2.50GHz running 24 physical cores). Both, the $x-$axis denoting the number of cores, and the $y-$axis denoting the calculation time in milliseconds are on the linear scale. Horizontally, the panels correspond to the different tree topologies, see also fig. 5.2. Vertically, the panels correspond to the different tree-sizes.

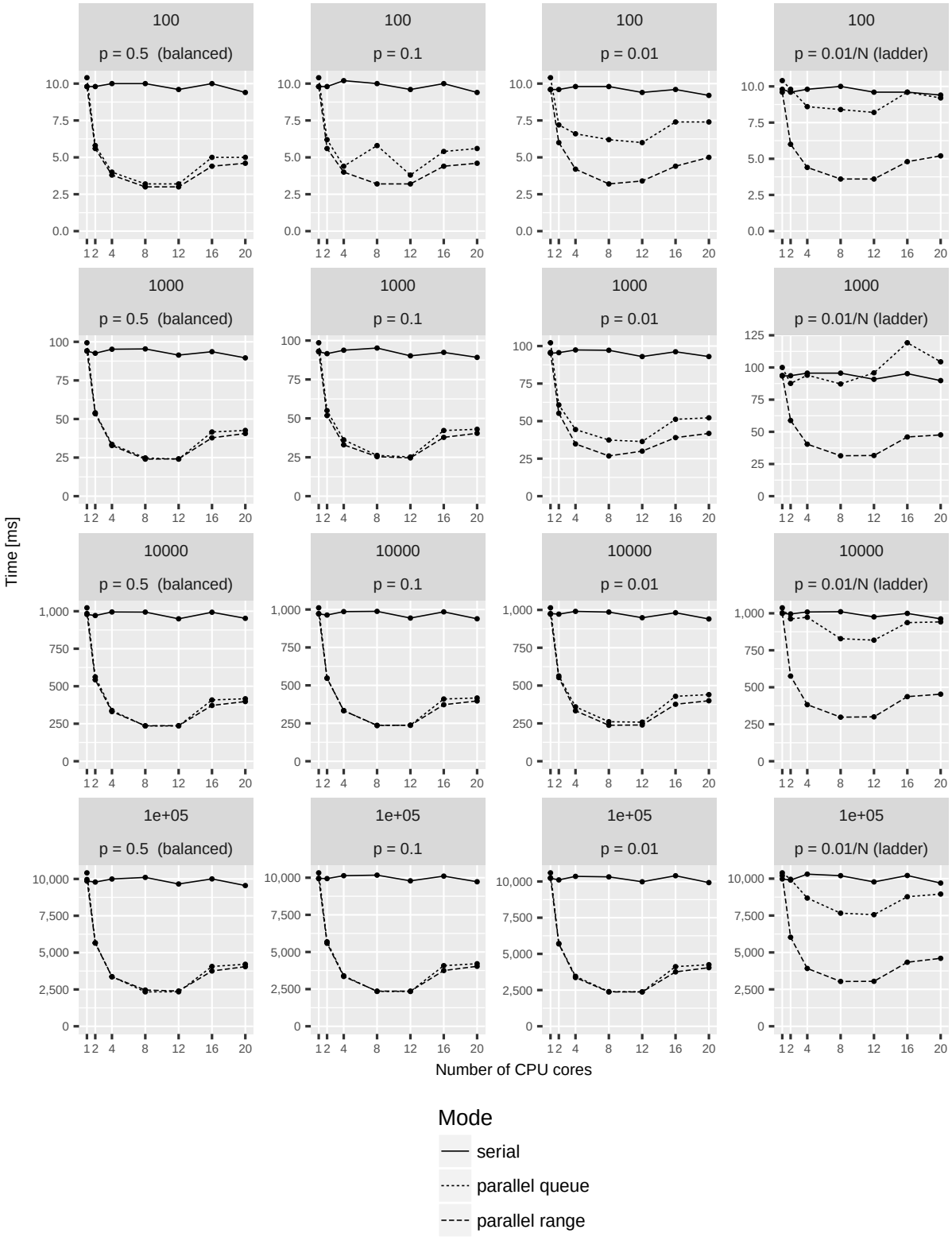


Figure S6: Likelihood calculation times for the multivariate POUMM implementation (`PCMBaseCpp`) with 8 traits on Euler cluster (a single shared memory node with Intel(R) Xeon(R) CPU E5-2680 v3 @ 2.50GHz running 24 physical cores). Both, the $x-$axis denoting the number of cores, and the $y-$axis denoting the calculation time in milliseconds are on the linear scale. Horizontally, the panels correspond to the different tree topologies, see also fig. 5.2. Vertically, the panels correspond to the different tree-sizes.

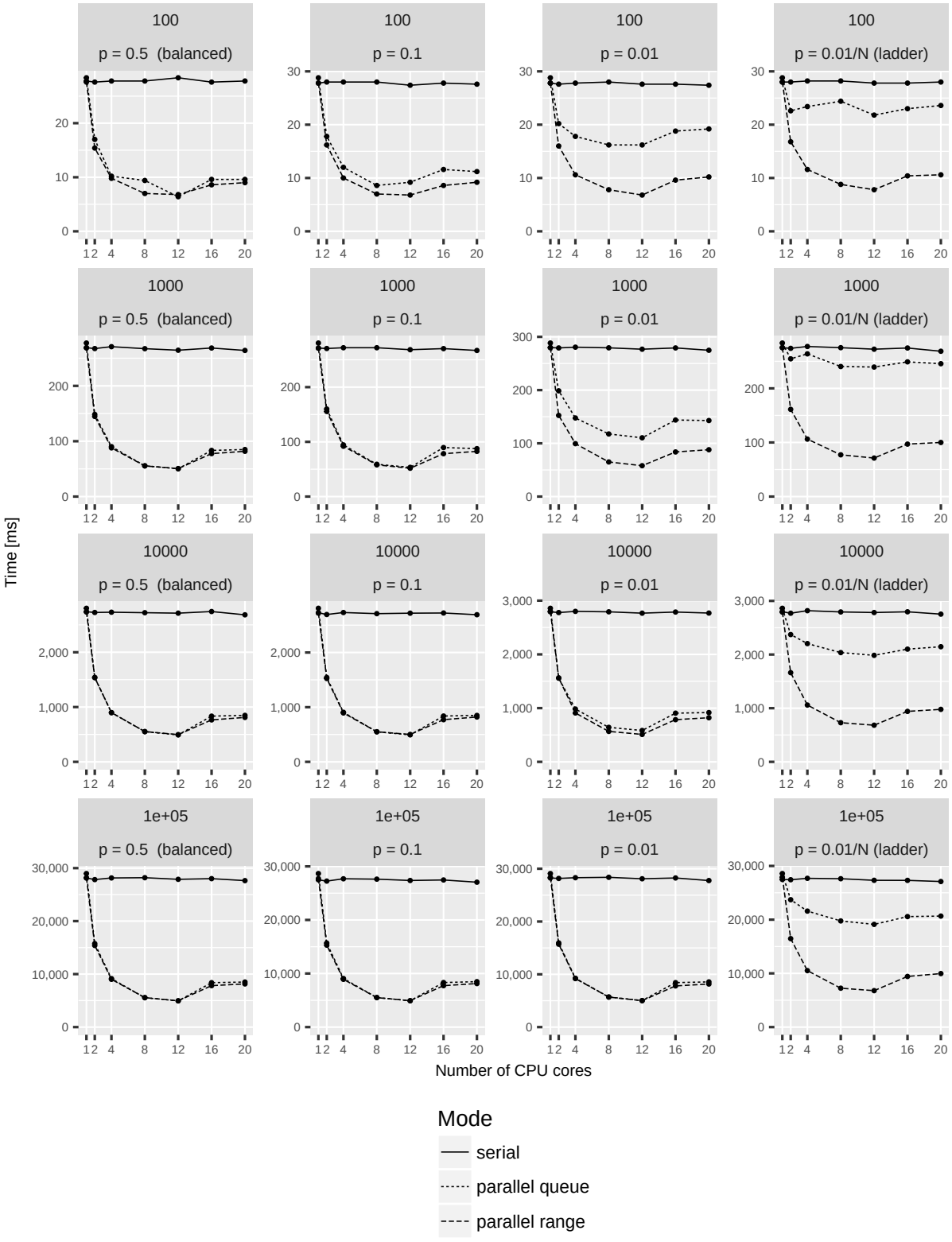


Figure S7: Likelihood calculation times for the multivariate POUMM implementation (`PCMBaseCpp`) with 16 traits on Euler cluster (a single shared memory node with Intel(R) Xeon(R) CPU E5-2680 v3 @ 2.50GHz running 24 physical cores). Both, the $x-$axis denoting the number of cores, and the $y-$axis denoting the calculation time in milliseconds are on the linear scale. Horizontally, the panels correspond to the different tree topologies, see also fig. 5.2. Vertically, the panels correspond to the different tree-sizes.

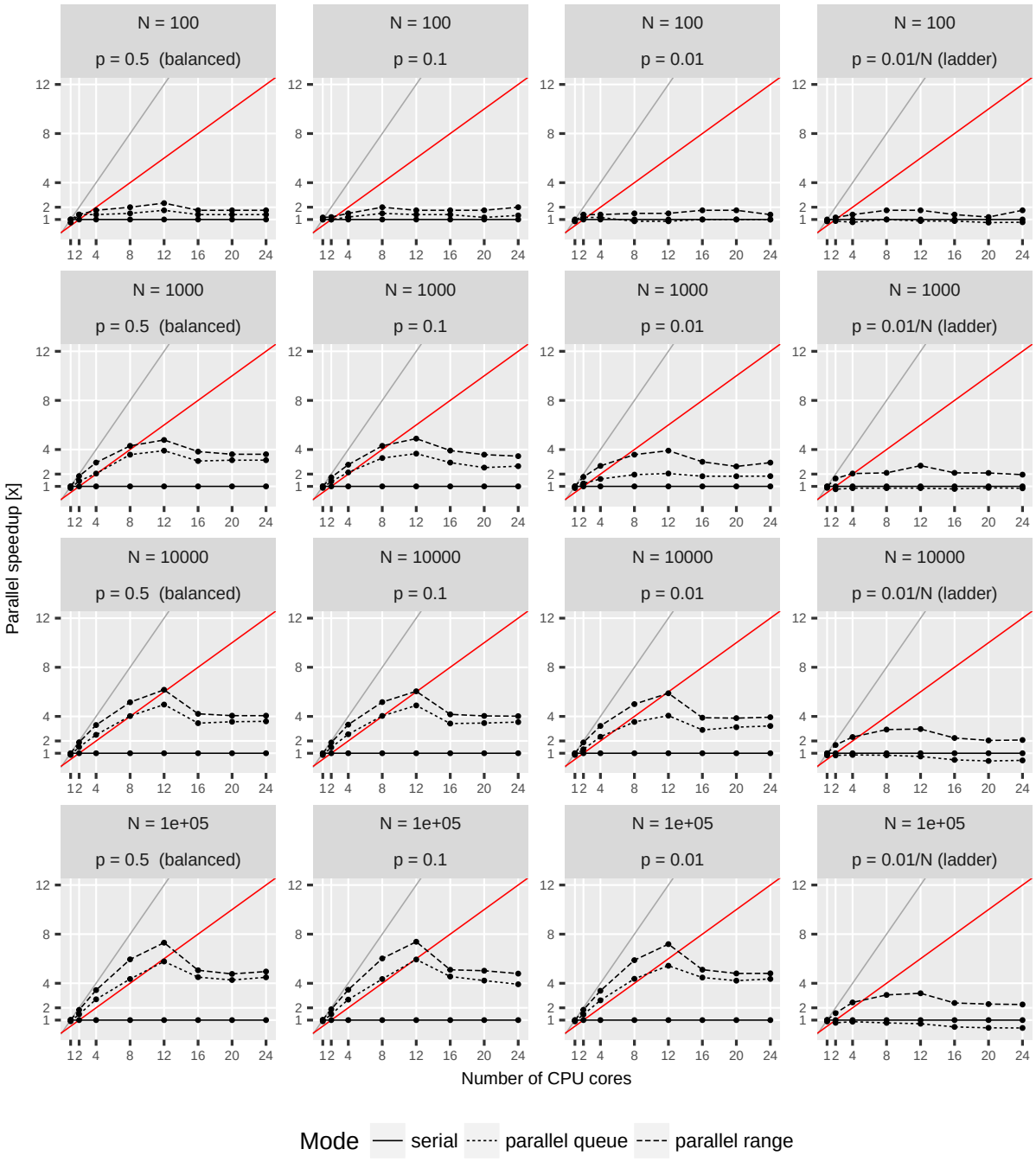


Figure S8: Parallel speed-up for the multivariate POUMM implementation (`PCMBaseCpp`) with 1 trait on Euler cluster. The grey and red lines denote, the expected speed-up at 100% and 50% parallel efficiency, respectively. Horizontally, the panels correspond to the different tree topologies. Vertically, the panels correspond to the different tree-sizes.

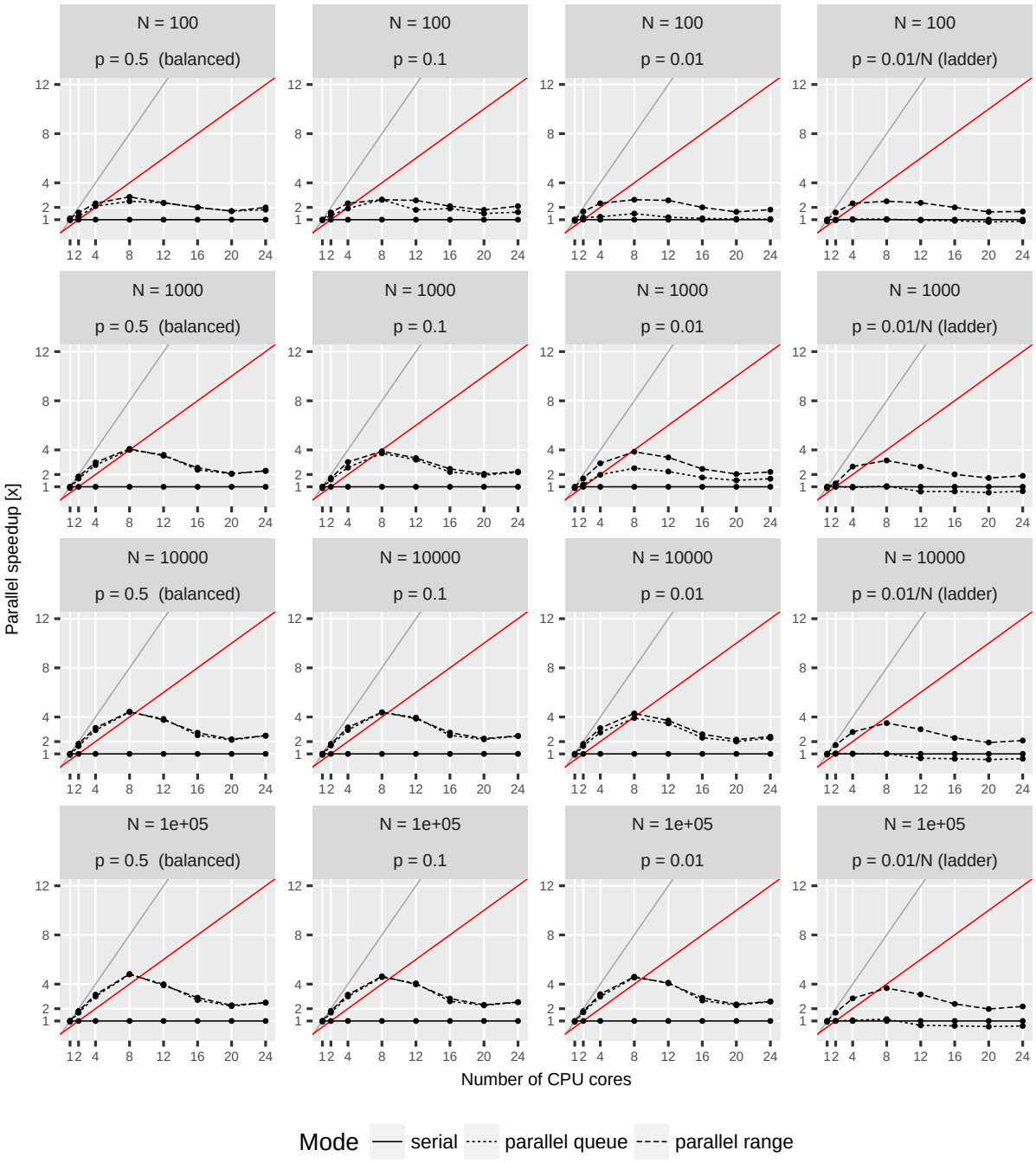


Figure S9: Parallel speed-up for the multivariate POUMM implementation (`PCMBaseCpp`) with 4 traits on Euler cluster. The grey and red lines denote, the expected speed-up at 100% and 50% parallel efficiency, respectively. Horizontally, the panels correspond to the different tree topologies. Vertically, the panels correspond to the different tree-sizes.

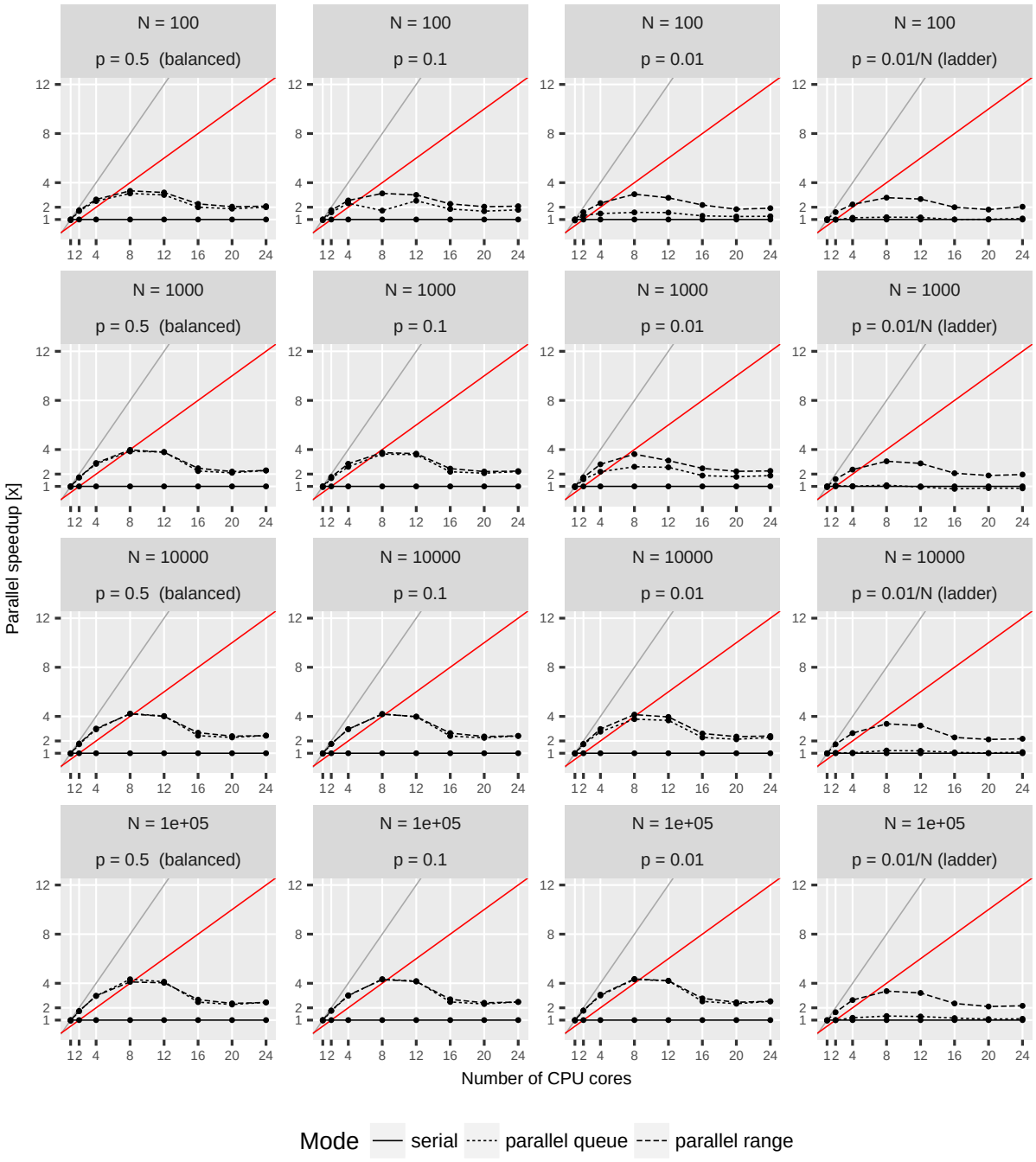


Figure S10: Parallel speed-up for the multivariate POUMM implementation (`PCMBaseCpp`) with 8 traits on Euler cluster. The grey and red lines denote, the expected speed-up at 100% and 50% parallel efficiency, respectively. Horizontally, the panels correspond to the different tree topologies. Vertically, the panels correspond to the different tree-sizes.

5.F COMBINED SPEED-UP FROM PARALLEL LIKELIHOOD CALCULATION AND ADAPTIVE METROPOLIS SAMPLING

We have used the POUMM package to estimate the heritability of set-point viral load in a data-set of 8,483 HIV patients. While the results of this analysis have been reported elsewhere (Mitov and Stadler, 2018), here, we briefly report the times and the quality statistics for the MCMC inference of the model with and without adaptive Metropolis sampling.

First, we ran the classical RWM Metropolis sampler with a default identity shape matrix for two MCMCs of ten million iterations on the above-mentioned hardware (2.3GHz Intel(R) Core i7 processor with 4 cores), using the fastest (range-based) parallel likelihood calculation. The total time for the two MCMCs was 3:18 hours. The run resulted in poor mixing and very low effective posterior sample size for most of the inferred parameters of the model (fig. S11a,b). The Gelman-Rubin statistic was greater than 1.1 for all parameters and the effective sample size was below 400 for all parameters, falling below 50 for α and σ.

Next, we ran the adaptive Metropolis sampler for two MCMCs of one million iterations. Adaptations has been enabled only for the first 100,000 iterations in each MCMC. The total runtime was 25 minutes. The two chains mixed very well and the effective sample size for all parameters exceeded 1200 (fig. S11c,d). The difference $|G.R. - 1|$ was below 0.01 for all parameters, proving that the MCMCs have converged to the same distribution, which is very likely the true posterior distribution for the model parameters.

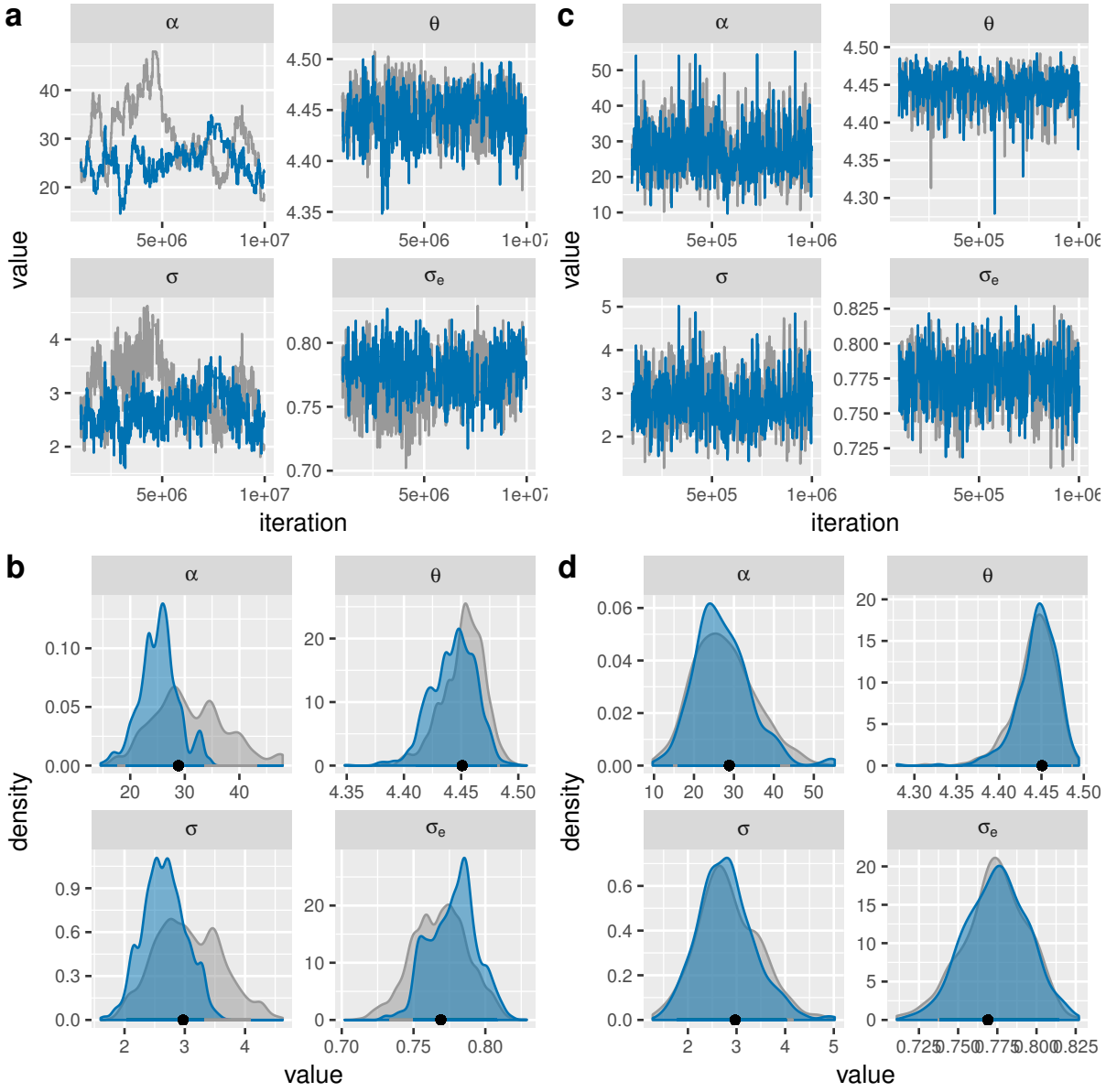


Figure S11: Sample trace- and density plots from a POUMM fit to a tree and virulence data 8483 HIV patients (Mitov and Stadler, 2018) a,b: no adaptation of the proposal shape matrix (ten million iterations); c,d: on-the-fly adaptation of the proposal shape matrix from the first 100,000 out of one million iterations. The colors correspond to the different chains.

6

FAST LIKELIHOOD EVALUATION FOR MULTIVARIATE PHYLOGENETIC COMPARATIVE METHODS

Manuscript to be submitted for peer review as

Venelin Mitov, Krzysztof Bartoszek, Georgios Asimomitis and Tanja Stadler (2018). Fast likelihood evaluation for multivariate phylogenetic comparative methods: the PCMBase R package *Journal of Theoretical Biology*.

This article represents the generalization of the quadratic polynomial representation for a single trait POUMM model discussed in Chapter 5 to multiple trait Gaussian phylogenetic models. The idea for this work came from my personal communication with Prof. Krzysztof Bartoszek. Krzysztof suggested that the integration over the unobserved internal nodes (Chapter 5, Appendix) could be generalized to multivariate Ornstein-Uhlenbeck processes. Later, we realized that the same holds for a much larger family, which we call $\mathcal{G}_{LInv}$. This paper describes the theoretical development and the design of an R-package called PCMBase. During the development of the PCMBase package I was assisted by Georgios Asimomitis – a master degree student in Computational Biology and Bioinformatics at ETH Zurich who did a 2-week lab rotation in the cEvo group. The PCMBase R-package will be the workhorse for the likelihood calculation of the mixed Gaussian phylogenetic models described in the next Chapter 7.

ABSTRACT

We introduce an 'R' package, '**PCMBase**', to rapidly calculate the likelihood for multivariate phylogenetic comparative methods. The package is not specific to particular models but offers the user the functionality to very easily implement a wide range of models where the transition along a branch is multivariate normal. We demonstrate the package's possibilities on the now standard, multitrait Ornstein–Uhlenbeck process as well as the novel multivariate punctuated equilibrium model. The package can handle trees of any type (e.g. ultrametric, nonultrametric, polytomies, e.t.c.), as well as complex cases of missing measurements or nonexisting traits for some of the species in the tree.

6.1 INTRODUCTION

Since Felsenstein (1985)'s work describing the independent contrasts algorithm, phylogenetic comparative methods (PCMs) have steadily been generalized with respect to available models and implementations of them. Following Felsenstein (1988)'s suggestion, Hansen (1997) described the Ornstein–Uhlenbeck (OU) process in the PCM setting. This led to the implementation of OU models in various packages such as '**ouch**' (Butler and King, 2004) or '**geiger**' (Harmon et al., 2008) to name a few, making it a standard model in the community alongside the Brownian motion (BM) process popularized in the community by Felsenstein (1985) but see also Edwards (1970) and Lande (1976). For species being characterized by multiple traits, the multivariate OU processes was introduced by 'R' packages such as '**ouch**', '**slouch**' (Hansen, Pienaar, and Orzack, 2008), '**mvSLOUCH**' (Bartoszek et al., 2012), '**mvMORPH**' (Clavel, Escarguel, and Merceron, 2015), '**Rphylopars**' (Goolsby, Bruggeman, and Ané, 2016), again, to name a few. At the core of these methods, the likelihood of the model parameters and tree for given trait data is evaluated, meaning the probability density of the tip trait values given the parameters and tree is calculated.

From a statistical point of view, the development of phylogenetic comparative methods goes in two directions. The first direction is development of model classes beyond simple stochastic processes, such as BM and OU, and the second direction is the development of efficient likelihood evaluation methods. Considering the first direction, we briefly mention three recent proposals. Manceau, Lambert, and Morlon (2016) show (with implementation in '**RPANDA**') that if one models the suite of traits by a linear stochastic differential equation (SDE, see the representation by Eq. (1) of Manceau, Lambert, and Morlon, 2016) whose drift matrix ("deterministic part" of the SDE) is piecewise constant with respect to the phylogeny, and diffusion matrix ("random part", sometimes referred to as "random drift part" in biological literature) does not depend on the trait, then the tip measurements are multivariate normal. The tip measurements' mean vector and covariance matrix can be found by integrating (backwards along the tree) an appropriate collection of ordinary differential equations (ODEs). Duchen et al. (2017) and Landis, Schraiber, and Liang (2012) went beyond the SDE world into Lévy process models. These are highly relevant from a biological point of view as they allow for jumps in the trait at random time instances. Hence, they hold promise for attacking the longstanding question of whether "evolution is gradual or punctuated?". Both approaches consider the transition densities, meaning the change of a trait between the start and the end of a branch, when quantifying trait evolution along a phylogeny. The third approach is to model the evolution of the traits' density in time with a partial differential equation (PDE Blomberg, 2017; Boucher et al., 2018). E.g. in the simplest standard Wiener process case the PDEs are $\frac{\partial}{\partial t} f_t(x) = \frac{1}{2} \frac{\partial^2}{\partial x^2} f_t(x)$ with boundary condition $f_0(x) = \delta_0(x)$, i.e.

Dirac δ at 0. This approach is convenient as it is next to impossible to analytically express the transition density.

The other direction is the development of efficient likelihood evaluation methods. Commonly, in PCMs, the model classes have the property that the joint distribution of the tip measurements is multivariate normal. Hence, there is a closed form for the likelihood—the multivariate normal density function, i.e. an algebraic expression in terms of the traits' mean vector and the traits' variance-covariance matrix ($\mathbf{V}$). Even though it is possible to obtain a conceptually simple equation, actually calculating the value of the likelihood is a computational challenge. If one has multiple correlated trait measurements per species, then the first step can be extremely involved, $\mathbf{V}$ can have a very complicated formula (cf. Eqs (A.1, B.3, B.7) of Bartoszek et al., 2012). As Freckleton (2012) points out "First, the matrix has to be generated in the first place. This requires allocating enough memory to hold all of the entries of $\mathbf{V}$ and then initiating one traversal (i.e. successively visiting all the nodes) of the phylogeny per pair of species sharing an ancestor to measure the shared path lengths. Second $\mathbf{V}$ has to be inverted at one point in the analysis.".

Hence, effort has been invested into reducing the memory and time complexity of the likelihood evaluation process. Inspired by Felsenstein (1973)'s approach, Freckleton (2012) proposed a linear way to obtain the likelihood for traits evolving as a Brownian motion. Freckleton (2012), further indicates that non–Brownian models can be quickly evaluated if one appropriately transforms the phylogeny. Then, Ho and Ané (2014a) proposed a general method that takes advantage of the so–called 3–point structure of the Brownian motion's between–species–between–traits variance–covariance matrix and obtain the likelihood in linear (w.r.t. the number of tips of the phylogeny) time, without having to construct in quadratic time the matrix $\mathbf{V}$. Similarly, calculating the likelihood for non–Brownian models (like the univariate Ornstein–Uhlenbeck process) can be done in linear time, as long as their $\mathbf{V}$ satisfies a generalized 3–point structure. Briefly, a covariance matrix satisfies the generalized 3–point structure if there exist diagonal matrices $\mathbf{D}_1$ and $\mathbf{D}_2$ such that $\mathbf{D}_1\mathbf{V}\mathbf{D}_2$ satisfies the 3–point structure Goolsby, Bruggeman, and Ané (2016) derives such a transformation to find the likelihood for traits under multivariate Ornstein–Uhlenbeck evolution in linear time. But in their implementation, only ultrametric trees and symmetric–positive–definite drift matrices are supported at the moment. For non–Gaussian models, a quasi–likelihood is defined and again the same approach (as long as the generalized 3–point structure holds) can be used (Ho and Ané, 2014a).

The speed–up for the Brownian motion's 3–point structure (or generalized 3-point structure) is based on the fact that the between–species–between–traits variance–covariance matrix has a nested structure. Therefore, appropriate linear algebra allows for rapid calculation of $\det(\mathbf{V})$ and quadratic forms like $\vec{x}\mathbf{V}^{-1}\vec{y}$ without the need to do the inversion $\mathbf{V}^{-1}$. (improved version of the Coppersmith–Winograd algorithm Le Gall, 2014).

Even though linear–time likelihood evaluation based on the 3–point structure is mathematically elegant, it is, due to the necessity of finding an appropriate transformation for non–Brownian motion, intrinsically complicated and may seem daunting for a non–algebraically oriented user or developer. FitzJohn (2012) indicated a probabilistically motivated way of quickly finding the likelihood (with implementation in the '**Diversitree**' 'R' package). He noticed (in the Supporting Information), same as Pybus et al. (2012), that one can traverse the tree and successively integrate out the internal nodes. FitzJohn (2012)'s description was focused around the BM and univariate OU processes on ultrametric trees. Furthermore, FitzJohn (2012) writes that he proved correctness of his method for a three tip phylogeny and then for larger trees checked numerically.

The presence of two different approaches, namely the 3–point structure method and the tree traversal method, to quickly calculating the likelihood combined with a number of independent implementations, each with some given set of conditions, can easily cause confusion. In fact, it seems that this led Slater (2014) to write in his Correction (due to "... errors arose from use of branch length rescaling under the Ornstein–Uhlenbeck process, which I here show to be inappropriate for non–ultrametric trees"), that "..., there is little, if any documentation in the literature or elsewhere highlighting that one of these approaches can be used while another cannot."

In this paper we attempt to assess the difficulties highlighted in the previous paragraph by proposing a fast method to obtain the likelihood which integrates over the internal node values. Our approach is appropriate for a large class of models, namely for all models where conditional on the ancestral trait, the descendant trait is normally distributed (however, we indicate in the Discussion that substantial relaxations of this are possible), the descendant's expectation depends linearly on the ancestor, and the variance does not depend on the ancestral value. In other words, we require that all transition densities along branches are Gaussian. From a mathematical point of view, we provide an inductive proof of FitzJohn (2012)'s claim of method correctness for multiple traits and all kinds of trees. Pybus et al. (2012) point out that for such a method to work, it is needed "to keep track of partial" means and precisions. Here, we propose a very general, computationally effective, and developer friendly way of doing this by recursively updating the polynomial representation of the multivariate normal density function. In order to use our approach for some new model, one has to be able to calculate the variance of the transition along the branch, the shift in the mean along a branch, and the linear dependency (i.e. a matrix) on the ancestral state. Thus, in our probabilistic approach, one needs to understand only the dynamics of a single branch (lineage), something that is usually present at the model formulation stage. For OU based models, these quantities can be analytically calculated and we provide an implementation. For other models, a developer will have to do the calculations themselves, but this should be significantly less involved than finding the transformation for the 3–point structure. In fact, for SDE based models, Manceau, Lambert, and Morlon (2016) provide a general ODE method (Eqs. S2 and S3) to obtain the conditional mean and variance. Furthermore, our method can naturally handle measurement error (intra–species variability), missing data, and punctuated components (jumps), and allows for changes in parameters at arbitrary points along the tree. It is further appropriate for non–ultrametric, binary and multifurcating trees. All of such specifications can be provided by the user. In no case is any tree transformation required.

As our method can handle any Gaussian transition, it encompasses a number of contemporary frameworks. In particular all OU type models (e.g. '**ouch**', '**slouch**', '**mvSLOUCH**', '**mvMORPH**') are covered by it. The '**RPANDA**' SDE framework (without interactions between lineages) is also covered as are current punctuated equilibrium models (OU along a branch with a normal jump Bartoszek, 2014; Bokma, 2002). To the best of our knowledge, our implementation handles the widest class of BM– and OU–based models on the widest set of phylogenetic trees, including non–ultrametric and non–binary trees.

It is important to stress here one point about the presented methodology and accompanying package. Our aim is not to provide a complete inference framework. Rather we provide an efficient way to evaluate the likelihood for a phylogenetic comparative data set given a user–defined model. The user can then on top of our package optimize over the parameter space to find the maximum likelihood estimates or perform a Bayesian analysis. In "Automatic Generation of Evolutionary Hypotheses using Mixed Gaussian Phylogenetic Models," we use the framework presented here to quantify the evolution of brain-body mass allometry in mammals.

The rest of the paper is organized as follows. In Section 6.2, we describe in detail our fast computational framework for phylogenetic comparative methods. In Section 6.3 we present the '**PCMBase**' 'R'–package. Then, in Section 6.4 we describe how one can handle issues such as missing values, measurement error, punctuated components, trees with polytomies, as well as sequentially sampled data (such as fossil data) leading to non–ultrametric trees. Next, in Section 6.5, we discuss the standard Ornstein–Uhlenbeck setup and describe examples of model classes that are already provided within our package. Two widely used models—the multivariate Brownian motion and multivariate Ornstein–Uhlenbeck processes and a novel model— a multivariate Ornstein–Uhlenbeck model with jumps are provided. It should be noted that even though we call the BM and OU standard PCM models, our implementation goes beyond what can be usually found in implementations: First, we allow for non-ultrametric trees [ok to put it here?]. Second, the only assumption that we make on the drift matrix (i.e. "deterministic part" of the SDE) is that it has to be eigendecomposable. This is in contrast to the assumption of this matrix being not only eigendecomposable but also non–singular (e.g. '**Rphylopars**', '**mvMORPH**', '**mvSLOUCH**'—but some exceptions to this are permitted). In section Section 6.6 we report a technical validation test of the likelihood calculations.

6.2 FAST PHYLOGENETIC COMPUTATIONAL FRAMEWORK

6.2.1 *Phylogenetic notation*

We assume that we are given a rooted phylogenetic tree $\mathbb{T}$ representing the ancestral relationship between N species associated with the tips of the tree (fig. 6.1). We denote the tips of the tree by the numbers $1, \ldots, N$, the internal nodes by the numbers $N+1, \ldots, M-1$ and the root-node by 0. For any internal node j, we denote by $Desc(j)$ the set of its direct descendants. We denote by $\mathbb{T}_j$ the subtree rooted at node j. We denote by t_j the known length of the branch in the tree leading to any tip or internal node j. By convention, we assume that time increases in the direction from the root to the tips of the tree, and t_j are positive scalars.

The object of all phylogenetic models discussed here will be a suite of k quantitative (real–valued) traits characterizing the N species. Associated with each tip, i, there is a real k–vector, $\vec{x}_i$, of measured values for the k traits. For some species, some trait measurements can be missing, reflecting two possible cases:

- the trait exists but was not measured for that species, denoted as '`NA`' (Not Available);
- the trait does not exist for that species denoted as '`NaN`' (Not a Number) (fig. 6.1).

We introduce algebraic notation that will hold for the rest of the paper. Scalars are denoted by lower case letters, e.g. f, vectors are indicated by the arrow notation, e.g. $\vec{\theta}$, while matrices are denoted as upper case bold letters, e.g. $\mathbf{H}$. An exception to this is $\mathbf{X}_j$, meaning the set of measurements at the tips descending from an internal node j of the tree.

6.2.2 *Phylogenetic models of continuous trait evolution*

We assume that the trait values measured at the tips of the tree result from a continuous state–space Markovian process evolving on top of the branching pattern in the tree. By this we mean that along any given branch we have a trajectory following the law of the process.

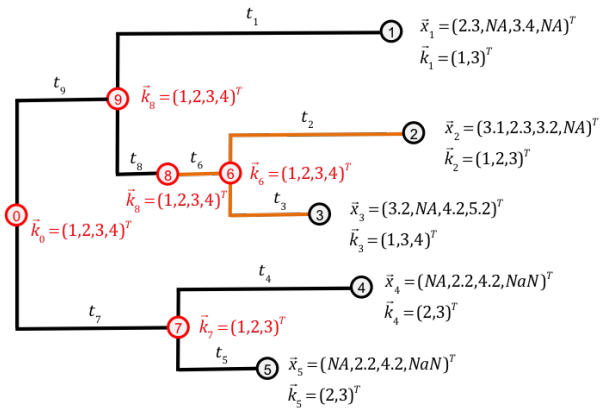


Figure 6.1: A phylogenetic tree with observations at the tips. Numbered circles in black indicate the tips with observed trait vectors $\vec{x}_1,\ldots,\vec{x}_{N=5}$. Missing measurements are denoted as '*NA*' (Not Available), while non-existing traits are denoted as '*NaN*' (Not a Number). Numbered circles in red indicate the root, 0, and the internal nodes $6,\ldots,9$, for which the trait vectors are unknown. The vectors, $\vec{k}_i$, denote the active coordinates for every node - for a tip-node these are all observed (neither '*NA*' nor '*NaN*') coordinates; for an internal node, these are all the coordinates denoting traits that exist (are not '*NaN*') for at least one of the tips descending from that node. The length of a branch leading to a tip or an internal node is known and denoted by t_i, $i=1,\ldots,9$. The change in branch color from black to orange at the internal node 8 denotes the change to a different evolutionary regime. It is assumed that such a regime change occurs simultaneously for all traits.

Then, at speciation, the process "splits" into two processes. Both processes inherit the last value of their parent process. After the branching points, there is no interaction between the processes. This entails that all the dependencies between the values at the tips come from the time between the origin of the tree and the most recent common ancestor for each pair of species. Exactly how this shared time of evolution is translated into a dependency depends on the assumed process. A widely used example of such trait process is the Ornstein–Uhlenbeck process illustrated on Fig. 6.2.

Such stochastic processes are used as models of continuous trait evolution at the macro-evolutionary time scale, that is, when the time-units are in the order of hundreds to thousands of generations. Previous works have studied the theoretical mapping of such processes to micro-evolutionary forces acting at the time-scale of single generations, e.g. random genetic drift and selection for reproduction (Hansen and Martins, 1996; Lande, 1976). Furhter in the text, we use the term "(trait evolutionary) model" to denote such kind of stochastic processes. We now turn to describing a family of models for which we will then provide an efficient way to calculate the likelihood of their parameters given the tree and the trait data observed at its tips.

6.2.3 *The $\mathcal{G}_{LInv}$ family of models*

The following definition specifies all requirements needed for a trait evolutionary model to be integrated within the fast computational framework:

Definition 1 (The $\mathcal{G}_{LInv}$ family)**.** *We say that a trait evolutionary model belongs to the $\mathcal{G}_{LInv}$ family if it satisfies the following*

1. *after branching the traits evolve independently in the two descending lineages,*

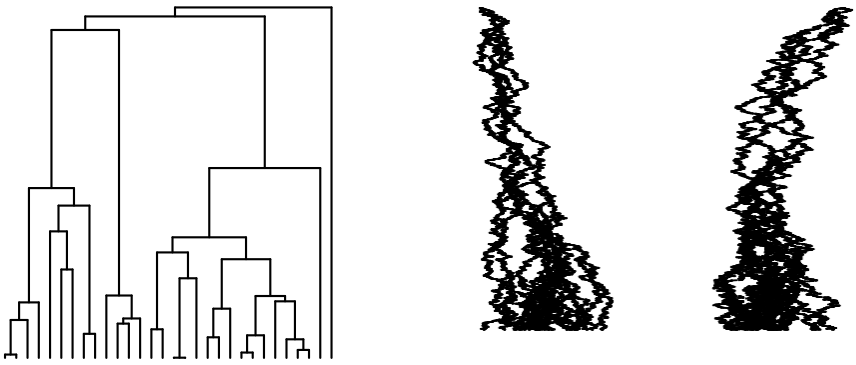


Figure 6.2: Simulation of a bivariate OU process on top of a pure birth tree with 30 tips. The two traits are displayed on separate panels. The tree was simulated using the '**TreeSim**' package (Stadler, 2009, 2011), its height is 3.201. The bivariate OU process was simulated using '**mvSLOUCH**' (Bartoszek et al., 2012) with parameters (matrices are represented by their rows) $\mathbf{H} = \{\{1, 0.25\}, \{0, 2\}\}$, $\mathbf{\Sigma}_x = \{\{0.5, 0.25\}, \{0, 0.5\}\}$, $\vec{\theta} = (1, -1)^T$ and $\vec{x}_0 = (0, 0)^T$.

2. *the distribution of the trait vector at time t, $\vec{x}(t)$, conditional on the trait vector at time $s < t$, $\vec{x}(s)$, is Gaussian with the mean and variance satisfying*

   (2.a) $\mathrm{E}\left[\vec{x}(t)|\vec{x}(s)\right] = \vec{\omega} + \mathbf{\Phi}\vec{x}(s)$
   (the expectation is a linearly function of the ancestral trait),

   (2.b) $\mathrm{Var}\left[\vec{x}(t)|\vec{x}(s)\right] = \mathbf{V}$
   (variance is invariant with respect to the ancestral trait),

   for some vector $\vec{\omega}$ and matrices $\mathbf{\Phi}$, $\mathbf{V}$ which may depend on s and t but do not depend on $\vec{x}(s)$.

Later, in section 6.5, we show that the $\mathcal{G}_{LInv}$ family contains many well-known contemporary models such as BM, multivariate OU (where all traits are OU or some are BMs), BM or OU with jumps (normally distributed). Now we derive an important property of the $\mathcal{G}_{LInv}$ family, namely, the equivalence between condition (2) in dfn. 1 and a mathematically convenient quadratic polynomial representation of the model transition density.

Theorem 2. *Let i be a tip or internal node and j be its parent node in $\mathbb{T}$. Let k_i and k_j be positive integers, and $\vec{x}_i \in \mathbb{R}^{k_i}$, $\vec{x}_j \in \mathbb{R}^{k_j}$ be the trait-vectors at nodes i and j. Assume that $\vec{x}_j$ is given and $\vec{x}_i$ is a random vector with non-zero support on the whole of $\mathbb{R}^{k_i}$. Then, the conditional distribution of $\vec{x}_i$ given $\vec{x}_j$ is a multivariate normal distribution satisfying the condition (2) in dfn. 1 iff there exists a symmetric negative–definite matrix $\mathbf{A}_i \in \mathbb{R}^{k_i \times k_i}$ and components $\vec{b}_i \in \mathbb{R}^{k_i}$, $\mathbf{C}_i \in \mathbb{R}^{k_j \times k_j}$, $\vec{d}_i \in \mathbb{R}^{k_j}$, $\mathbf{E}_i \in \mathbb{R}^{k_j \times k_i}$, $f_i \in \mathbb{R}$, all of which independent of $\vec{x}_j$ and satisfying the equation*

$$pdf(\vec{x}_i|\vec{x}_j) = \exp(\vec{x}_i^T \mathbf{A}_i \vec{x}_i + \vec{x}_i^T \vec{b}_i + \vec{x}_j^T \mathbf{C}_i \vec{x}_j + \vec{x}_j^T \vec{d}_i + \vec{x}_j^T \mathbf{E}_i \vec{x}_i + f_i), \tag{6.1}$$

Furthermore, the components of Eq. (6.1) *must satisfy the constraints*

$$\begin{aligned} \mathbf{C}_i &= \mathbf{E}_i \mathbf{A}_i^{-T} \mathbf{E}_i^T, \\ \vec{d}_i &= 2\mathbf{E}_i \mathbf{A}_i^{-1} \vec{b}_i, \\ f_i &= 4 b_i^T b_i - \tfrac{k_i}{2} \log \pi - \tfrac{1}{2} \log |\mathbf{A}_i|. \end{aligned} \tag{6.2}$$

Proof.
($\Longrightarrow$)
We denote the elements $\vec{\omega}$, $\mathbf{\Phi}$ and $\mathbf{V}$ (dfn. 1) specific for node i, by $\vec{\omega}_i$, $\mathbf{\Phi}_i$ and $\mathbf{V}_i$. Substituting $\vec{\omega}_i + \mathbf{\Phi}_i \vec{x}_j$ and $\mathbf{V}_i$ for the mean and variance in the formula for the multivariate Gaussian pdf, we obtain:

$$pdf(\vec{x}_i|\vec{x}_j) = \exp\left(-\frac{1}{2} \left(\vec{x}_i - \left(\vec{\omega}_i + \mathbf{\Phi}_i \vec{x}_j\right)\right)^T \mathbf{V}_i^{-1} \left(\vec{x}_i - \left(\vec{\omega}_i + \mathbf{\Phi}_i \vec{x}_j\right)\right) - \frac{k_i}{2} \log(2\pi) - \frac{1}{2} \log |\mathbf{V}_i| \right) \tag{6.3}$$

Equation (6.3) can be rewritten as

$$\begin{aligned} &\exp\Big(-\tfrac{1}{2} \vec{x}_i^T \mathbf{V}_i^{-1} \vec{x}_i + \vec{x}_i^T \mathbf{V}_i^{-1} \vec{\omega}_i + \vec{x}_j^T \mathbf{\Phi}_i^T \mathbf{V}_i^{-1} \vec{x}_i - \tfrac{1}{2} \vec{\omega}_i^T \mathbf{V}_i^{-1} \vec{\omega}_i - \vec{x}_j^T \mathbf{\Phi}_i^T \mathbf{V}_i^{-1} \vec{\omega}_i - \tfrac{1}{2} \vec{x}_j^T \mathbf{\Phi}_i^T \mathbf{V}_i^{-1} \mathbf{\Phi}_i \vec{x}_j \\ &- \tfrac{k_i}{2} \log(2\pi) - \tfrac{1}{2} \log |\mathbf{V}_i| \Big). \end{aligned} \tag{6.4}$$

We can see the correspondence with the parametrization of Eq. (6.1)

$$\begin{aligned} \mathbf{A}_i &= -\tfrac{1}{2} \mathbf{V}_i^{-1} \in \mathbb{R}^{k_i \times k_i} \\ \vec{b}_i &= \mathbf{V}_i^{-1} \vec{\omega}_i \in \mathbb{R}^{k_i} \\ \mathbf{C}_i &= -\tfrac{1}{2} \mathbf{\Phi}_i^T \mathbf{V}_i^{-1} \mathbf{\Phi}_i \in \mathbb{R}^{k_j \times k_j} \\ \vec{d}_i &= -\mathbf{\Phi}_i^T \mathbf{V}_i^{-1} \vec{\omega}_i \in \mathbb{R}^{k_j} \\ \mathbf{E}_i &= \mathbf{\Phi}_i^T \mathbf{V}_i^{-1} \in \mathbb{R}^{k_j \times k_i} \\ f_i &= -\tfrac{1}{2} \vec{\omega}_i^T \mathbf{V}_i^{-1} \vec{\omega}_i - \tfrac{k_i}{2} \log(2\pi) - \tfrac{1}{2} \log |\mathbf{V}_i| \in \mathbb{R}. \end{aligned} \tag{6.5}$$

We notice that $\mathbf{V}_i^{-1}$ is symmetric positive–definite, as the inverse of $\mathbf{V}_i$, which is symmetric positive–definite by definition. This implies that the term $\mathbf{A}_i$ is symmetric negative–definite. Further, by dfn. 1, $\vec{\omega}_i$, $\mathbf{\Phi}_i$ and $\mathbf{V}_i$ are independent of $\vec{x}_j$. Hence, $\mathbf{A}_i$, $\vec{b}_i$, $\mathbf{C}_i$, $\vec{d}_i$, $\mathbf{E}_i$, f_i are also independent with respect to $\vec{x}_j$. Validating Eq. (6.2) is a matter of simple algebraic conversion.
($\Longleftarrow$)
We rearrange the terms on the right-hand side of Eq. (6.1) as follows

$$\begin{aligned} &pdf(\vec{x}_i|\vec{x}_j) = \exp\left(\vec{x}_i^T \mathbf{A}_i \vec{x}_i - 2\vec{x}_i^T \mathbf{A}_i \left((-\tfrac{1}{2} \mathbf{A}_i^{-1}) \left(\vec{b}_i + \mathbf{E}_i^T \vec{x}_j \right) \right) + \left(\vec{x}_j^T \mathbf{C}_i \vec{x}_j + \vec{x}_j^T \vec{d}_i + f_i \right) \right) \\ &= \exp\Bigg(\left(\vec{x}_i + \tfrac{1}{2} \mathbf{A}_i^{-1} \left(\vec{b}_i + \mathbf{E}_i^T \vec{x}_j \right) \right)^T \mathbf{A}_i \left(\vec{x}_i + \tfrac{1}{2} \mathbf{A}_i^{-1} \left(\vec{b}_i + \mathbf{E}_i^T \vec{x}_j \right) \right) - \tfrac{1}{4} \left(\vec{b}_i + \mathbf{E}_i^T \vec{x}_j \right)^T \mathbf{A}_i^{-1} \left(\vec{b}_i + \mathbf{E}_i^T \vec{x}_j \right) \\ &+ \left(\vec{x}_j^T \mathbf{C}_i \vec{x}_j + \vec{x}_j^T \vec{d}_i + f_i \right) \Bigg). \end{aligned}$$

As the above is by definition a density on $\mathbb{R}^{k_i}$, we have

$$
\begin{aligned}
1 &= \int_{\mathbb{R}^{k_i}} \exp\left(-\tfrac{1}{2}\left(\vec{x}_i + \tfrac{1}{2}\mathbf{A}_i^{-1}\left(\vec{b}_i + \mathbf{E}_i^T\vec{x}_j\right)\right)^T(-2\mathbf{A}_i)\left(\vec{x}_i + \tfrac{1}{2}\mathbf{A}_i^{-1}\left(\vec{b}_i + \mathbf{E}_i^T\vec{x}_j\right)\right)\right)\mathrm{d}\vec{x}_i \\
&\cdot \exp\left(-\tfrac{1}{4}\left(\vec{b}_i + \mathbf{E}_i^T\vec{x}_j\right)^T\mathbf{A}_i^{-1}\left(\vec{b}_i + \mathbf{E}_i^T\vec{x}_j\right) + \left(\vec{x}_j^T\mathbf{C}_i\vec{x}_j + \vec{x}_j^T\vec{d}_i + f_i\right)\right) \\
&= \exp\left(\tfrac{k_i}{2}\log(2\pi) + \tfrac{1}{2}\log|(-2)\mathbf{A}_i|\right) \times \exp\left(-\tfrac{1}{4}\left(\vec{b}_i + \mathbf{E}_i^T\vec{x}_j\right)^T\mathbf{A}_i^{-1}\left(\vec{b}_i + \mathbf{E}_i^T\vec{x}_j\right) + \left(\vec{x}_j^T\mathbf{C}_i\vec{x}_j + \vec{x}_j^T\vec{d}_i + f_i\right)\right) \\
&= \exp\left(\vec{x}_j^T\left(\mathbf{C}_i - \tfrac{1}{4}\mathbf{E}_i\mathbf{A}_i^{-1}\mathbf{E}_i^T\right)\vec{x}_j + \vec{x}_j^T\left(\vec{d}_i - \tfrac{1}{2}\mathbf{E}_i\mathbf{A}_i^{-1}\vec{b}_i\right) + f_i + \tfrac{k_i}{2}\log(2\pi) + \tfrac{1}{2}\log|(-2)\mathbf{A}_i| - \tfrac{1}{4}\vec{b}_i^T\mathbf{A}_i^{-1}\vec{b}_i\right).
\end{aligned}
\tag{6.6}
$$

By definition, $\mathbf{A}_i$, $\vec{b}_i$, $\mathbf{C}_i$, $\vec{d}_i$, $\mathbf{E}_i$, f_i are independent with respect to $\vec{x}_j$. Therefore, eq. 6.6 has to hold for all $\vec{x}_j$. This implies $\mathbf{C}_i = \frac{1}{4}\mathbf{E}_i\mathbf{A}_i^{-1}\mathbf{E}_i^T$, $\vec{d}_i = \frac{1}{2}\mathbf{E}_i\mathbf{A}_i^{-1}\vec{b}_i$ and $f_i = -\frac{k_i}{2}\log(2\pi) - \frac{1}{2}\log|(-2)\mathbf{A}_i| + \frac{1}{4}\vec{b}_i^T\mathbf{A}_i^{-1}\vec{b}_i$. With that, we obtained the constraints of Eq. (6.2). Next, we define $\mathbf{V}_i := (-\frac{1}{2})\mathbf{A}_i^{-1}$, $\vec{\omega}_i := (-\frac{1}{2})\mathbf{A}_i^{-1}\vec{b}_i$ and $\mathbf{\Phi}_i := (-\frac{1}{2})\mathbf{A}_i^{-1}\mathbf{E}_i^T$. Since $\mathbf{A}_i$ is symmetric negative–definite, $\mathbf{V}_i$ is symmetric positive–definite. Combining the above three definitions with Eq. (6.2) and expressing $\mathbf{A}_i$, $\vec{b}_i$, $\mathbf{C}_i$, $\vec{d}_i$, $\mathbf{E}_i$, f_i in terms of $\vec{\omega}_i$, $\mathbf{\Phi}_i$ and $\mathbf{V}_i$, we obtain again eq. (6.5). Then, we can follow the equivalences in backward direction (eqs. 6.5→6.4→6.3) to prove that the pdf defined in eq. 6.1 is equivalent to the Gaussian pdf defined in terms of $\vec{\omega}_i$, $\mathbf{\Phi}_i$ and $\mathbf{V}_i$, eq. 6.3. □

6.2.4 *Analytical integration over the internal nodes*

The representation of Eq. (6.1) allows for linear (in terms of tips) calculation of the likelihood over a given phylogeny. This follows from the next theorem.

Theorem 3. *With the representation of Eq.* (6.1)*, for the root or an internal node j, there exists a $k_j \times k_j$ matrix $\mathbf{L}_j$, a k_j–vector $\vec{m}_j$ and a scalar r_j, such that the likelihood for $\mathbf{X}_j$ conditioned on $\vec{x}_j \in \mathbb{R}^{k_j}$ and $\mathbb{T}_j$ (the subtree with node j as its root) is expressed as:*

$$
pdf(\mathbf{X}_j|\vec{x}_j, \mathbb{T}_j) = \exp\left(\vec{x}_j^T\mathbf{L}_j\vec{x}_j + \vec{x}_j^T\vec{m}_j + r_j\right). \tag{6.7}
$$

The parameters $\mathbf{L}_j$, $\vec{m}_j$, r_j are functions of the model parameters Θ, the observed data $\mathbf{X}_j$, and the tree $\mathbb{T}_j$ both in terms of topology and branch lengths, namely, equations 6.10, 6.11, and 6.12.

Proof. We consider the factorization of the conditional likelihood at any internal or root node j. Splitting $Desc(j)$, i.e. the set of nodes descending from node j, into tips and non–tips, denoted as $Desc(j) \cap \{1, ..., N\}$ and $Desc(j) \setminus \{1, ..., N\}$, we can write:

$$
pdf(\mathbf{X}_j|\vec{x}_j, \mathbb{T}_j) = \left(\prod_{i \in Desc(j)\cap\{1,...,N\}} pdf(\vec{x}_i|\vec{x}_j, t_i)\right) \times \left(\prod_{i \in Desc(j)\setminus\{1,...,N\}} \int_{\mathbb{R}_j^k} pdf(\vec{x}_i|\vec{x}_j, t_i) \times pdf(\mathbf{X}_i|\vec{x}_i, \mathbb{T}_i)\mathrm{d}\vec{x}_i\right). \tag{6.8}
$$

If all descendants of j are tips (e.g. nodes 6 and 7 on Fig. 6.1), then, according to Eq. (6.1)

$$\begin{aligned} pdf(\mathbf{X}_j|\vec{x}_j,\mathbb{T}_j) &= \prod_{i\in Desc(j)} pdf(\vec{x}_i|\vec{x}_j,t_i) \\ &= \exp\left(\sum_{i\in Desc(j)} \vec{x}_i^T\mathbf{A}_i\vec{x}_i + \vec{x}_i^T\vec{b}_i + \vec{x}_j^T\mathbf{C}_i\vec{x}_j + \vec{x}_j^T\vec{d}_i + \vec{x}_j^T\mathbf{E}_i\vec{x}_i + f_i\right) \\ &= \exp\left(\vec{x}_j^T(\sum_{i\in Desc(j)}\mathbf{C}_i)\vec{x}_j + \vec{x}_j^T(\sum_{i\in Desc(j)}\vec{d}_i + \mathbf{E}_i\vec{x}_i) + \sum_{i\in Desc(j)}\vec{x}_i^T\mathbf{A}_i\vec{x}_i + \vec{x}_i^T\vec{b}_i + f_i\right) \end{aligned} \tag{6.9}$$

Then, to obtain the representation from Eq. (6.7), we denote:

$$\begin{aligned} \mathbf{L}_j &= \sum_{i\in Desc(j)} \mathbf{C}_i \\ \vec{m}_j &= \sum_{i\in Desc(j)} \vec{d}_i + \mathbf{E}_i\vec{x}_i \\ r_j &= \sum_{i\in Desc(j)} \vec{x}_i^T\mathbf{A}_i\vec{x}_i + \vec{x}_i^T\vec{b}_i + f_i \end{aligned} \tag{6.10}$$

If not all of $Desc(j)$ are tips, then, for the descendants which are tips, we define:

$$\begin{aligned} \mathbf{L}_j^{tips} &= \sum_{i\in Desc(j)\cap\{1,\ldots,N\}} \mathbf{C}_i \\ \vec{m}_j^{tips} &= \sum_{i\in Desc(j)\cap\{1,\ldots,N\}} \vec{d}_i + \mathbf{E}_i\vec{x}_i \\ r_j^{tips} &= \sum_{i\in Desc(j)\cap\{1,\ldots,N\}} \vec{x}_i^T\mathbf{A}_i\vec{x}_i + \vec{x}_i^T\vec{b}_i + f_i \end{aligned} \tag{6.11}$$

Following mathematical induction and the reasoning behind Eqs. (6.9) and (6.10), for each $i \in Desc(j) \setminus \{1,\ldots,N\}$ there exists a $k_i \times k_i$ matrix $\mathbf{L}_i$, a k_i–vector $\vec{m}_i$ and a scalar r_i such that $pdf(\mathbf{X}_i|\vec{x}_i,\mathbb{T}_i) = \exp(\vec{x}_i^T\mathbf{L}_i\vec{x}_i + \vec{x}_i^T\vec{m}_i + r_i)$. To be more precise the initial step of the induction is what we proved above, the quadratic polynomial representation for branches leading to tips. Then, the induction hypothesis is that for an internal node j, the statement of the theorem has been proven for all subtrees $\mathbb{T}_i$, such that $i \in Desc(j)$. Now in inductive step using Eq. (6.1) and the induction hypothesis, we can write the integral in Eq. (6.8) as

$$
\begin{aligned}
&\int_{\mathbb{R}^{k_i}} pdf(\vec{x}_i|\vec{x}_j, t_i) \times pdf(\mathbf{X}_i|\vec{x}_i, \mathbb{T}_i)\mathrm{d}\vec{x}_i \\
&= \int_{\mathbb{R}^{k_i}} \exp\left(\vec{x}_i^T \mathbf{A}_i \vec{x}_i + \vec{x}_i^T \vec{b}_i + \vec{x}_j^T \mathbf{C}_i \vec{x}_j + \vec{x}_j^T \vec{d}_i + \vec{x}_j^T \mathbf{E}_i \vec{x}_i + f_i + \vec{x}_i^T \mathbf{L}_i \vec{x}_i + \vec{x}_i^T \vec{m}_i + r_i\right) \mathrm{d}\vec{x}_i \\
&= \exp\left(\vec{x}_j^T \mathbf{C}_i \vec{x}_j + \vec{x}_j^T \vec{d}_i + f_i + r_i\right) \times \boxed{\int_{\mathbb{R}^{k_i}} \exp\left(\vec{x}_i^T (\mathbf{A}_i + \mathbf{L}_i)\vec{x}_i + \vec{x}_i^T (\vec{b}_i + \vec{m}_i + \mathbf{E}_i^T \vec{x}_j)\right) \mathrm{d}\vec{x}_i} \\
&\overset{\bigstar}{=} \exp\left(\vec{x}_j^T \mathbf{C}_i \vec{x}_j + \vec{x}_j^T \vec{d}_i + f_i + r_i\right) \left(\sqrt{2\pi}\right)^{k_i} \boxed{\left(\sqrt{|(-2)\,(\mathbf{A}_i + \mathbf{L}_i)\,|}\right)^{-1}} \\
&\qquad \times \exp\left(-(1/4)\left(\vec{b}_i + \vec{m}_i + \mathbf{E}_i^T \vec{x}_j\right)^T (\mathbf{A}_i + \mathbf{L}_i)^{-1} \left(\vec{b}_i + \vec{m}_i + \mathbf{E}_i^T \vec{x}_j\right)\right)
\end{aligned}
$$

$$
\begin{aligned}
&= \exp\left(\vec{x}_j^T \mathbf{C}_i \vec{x}_j + \vec{x}_j^T \vec{d}_i + f_i + r_i\right) \left(\sqrt{2\pi}\right)^{k_i} \left(\sqrt{|(-2)\,(\mathbf{A}_i + \mathbf{L}_i)\,|}\right)^{-1} \\
&\times \exp\left(-(1/4)\left(\vec{b}_i + \vec{m}_i\right)^T (\mathbf{A}_i + \mathbf{L}_i)^{-1} \left(\vec{b}_i + \vec{m}_i\right) - (1/2)\vec{x}_j^T \mathbf{E}_i \,(\mathbf{A}_i + \mathbf{L}_i)^{-1} \left(\vec{b}_i + \vec{m}_i\right)\right. \\
&\left. -(1/4)\vec{x}_j^T \mathbf{E}_i \,(\mathbf{A}_i + \mathbf{L}_i)^{-1} \mathbf{E}_i^T \vec{x}_j\right) \\
&= \exp\left(\vec{x}_j^T \left(\mathbf{C}_i - (1/4)\mathbf{E}_i \,(\mathbf{A}_i + \mathbf{L}_i)^{-1} \mathbf{E}_i^T\right) \vec{x}_j + \vec{x}_j^T \left(\vec{d}_i - (1/2)\mathbf{E}_i \,(\mathbf{A}_i + \mathbf{L}_i)^{-1} \left(\vec{b}_i + \vec{m}_i\right)\right)\right. \\
&\left. + f_i + r_i + (k_i/2)\log(2\pi) - (1/2)\log(|(-2)\,(\mathbf{A}_i + \mathbf{L}_i)\,|) - (1/4)\left(\vec{b}_i + \vec{m}_i\right)^T (\mathbf{A}_i + \mathbf{L}_i)^{-1} \left(\vec{b}_i + \vec{m}_i\right)\right)
\end{aligned}
$$

We can then see that for a non–tip node we can define

$$
\begin{aligned}
\mathbf{L}_j^{non-tips} &= \sum_{i \in Desc(j) \setminus \{1,\ldots,N\}} \left(\mathbf{C}_i - (1/4)\mathbf{E}_i \,(\mathbf{A}_i + \mathbf{L}_i)^{-1} \mathbf{E}_i^T\right) \\
\vec{m}_j^{non-tips} &= \sum_{i \in Desc(j) \setminus \{1,\ldots,N\}} \left(\vec{d}_i - (1/2)\mathbf{E}_i \,(\mathbf{A}_i + \mathbf{L}_i)^{-1} \left(\vec{b}_i + \vec{m}_i\right)\right) \\
r_j^{non-tips} &= \sum_{i \in Desc(j) \setminus \{1,\ldots,N\}} (f_i + r_i + (k_i/2)\log(2\pi) - (1/2)\log(|(-2)\,(\mathbf{A}_i + \mathbf{L}_i)\,|) \\
&\quad -(1/4)\left(\vec{b}_i + \vec{m}_i\right)^T (\mathbf{A}_i + \mathbf{L}_i)^{-1} \left(\vec{b}_i + \vec{m}_i\right)\Big).
\end{aligned}
\tag{6.12}
$$

The representation of $\mathbf{L}_j^{non-tips}$, $\vec{m}_j^{non-tips}$ and $r_j^{non-tips}$ in Eq. (6.12) immediately entails the existence of the $\mathbf{L}_j$, $\vec{m}_j$ and r_j elements in Eq. (6.7) for internal or root nodes j, hence we obtain the claimed polynomial form in the inductive step and in consequence the theorem. □

The inductive proof of Thm. 3 defines a pruning–wise procedure for calculating $\mathbf{L}_0$, $\vec{m}_0$ and r_0 (we remind that 0 stands for the root of the tree). In order to calculate the likelihood of the tree conditioned on $\vec{x}_0$, we use Thm 3 with j being the root node. In order to be able to calculate the full likelihood, it now only remains to specify how to deal with the unknown trait value at the root of the tree, $\vec{x}_0$, i.e. the ancestral state. This is an implementation detail up to the user. Our implementation of the various models provided (sections 6.3 and 6.5) with the '**PCMbase**' package allow for maximizing the polynomial with respect to $\vec{x}_0$ or

for treating it as a free parameter (like the elements of the parameter set Θ) that the user provides.

6.3 THE '**pcmbase**' 'r' PACKAGE

The '**PCMBase**' package takes advantage of the fact that the quadratic polynomial representation of the likelihood function is valid for all models in the $\mathcal{G}_{LInv}$ family. Hence, once the analytical integration over the internal nodes has been implemented, the addition of a new $\mathcal{G}_{LInv}$ model to the framework boils down to defining the transition density in terms of the functions $\vec{\omega}$, $\mathbf{\Phi}$ and $\mathbf{V}$ (Def. 1). '**PCMBase**' implements this idea, based on the concept of inheritance between programming modules: Eqs. (6.5), (6.10), (6.11), (6.12)) are implemented in a base module called "GaussianPCM", which is abstract with respect to $\vec{\omega}$, $\mathbf{\Phi}$ and $\mathbf{V}$ (Fig. 6.3). These functions are provided in inheriting modules definable for each $\mathcal{G}_{LInv}$ model. This hierarchical design is shown on Fig. 2.

6.3.1 *Extending '**PCMBase**'*

Extending the '**PCMBase**' functionality can be achieved in two ways:

1. **Adding a new model**. It is possible to write a new module inheriting from the module "GaussianPCM" and implementing its own version of the functions $\vec{\omega}$, $\mathbf{\Phi}$ and $\mathbf{V}$;
2. **Adding a parametrization**. It is possible to restrict or apply a transformation to some of the parameters of an already defined model (Fig. 6.3).

6.3.2 *Using the package*

Figure 6.4 shows the runtime objects and use-cases currently implemented in the '**PCMBase**' package. Once the modules for the models of interest have been implemented, the '**PCMBase**' package can be used to:

- Creating a model object. The end-user function for creating a model object is '`PCM()`'. A model object represents an S3 object, that is, a named list with members corresponding to the model parameters, such as '`H`', '`Sigma_x`' and '`Sigmae_x`', and a class attribute equalling the model type, e.g. '`BM`' or '`OU`'.
- Simulating the evolution of a set of continuous traits along a tree, according to a model. The user level function for trait simulation is '`PCMSim()`'. Based on the S3 class of its model argument '`PCMSim()`' invokes an appropriate specification of the S3 generic function '`PCMCond()`', which creates a random sampler from the trait distribution at the end of a branch, given the model, the branch length and the trait values at the beginning of the branch.
- Calculating the (log–)likelihood of a model, given a tree and trait values at its tips. The user level function for likelihood calculation is '`PCMLik()`'. This function is implemented in the "GaussianPCM" module and inherited by all of its daughter modules. The calculation proceeds in four steps:
    1. Initially, the model-specific functions $\vec{\omega}$, $\mathbf{\Phi}$ and $\mathbf{V}$ are calculated based on the model parameters Θ and the branch lengths t_i (note that this operation does not need the trait values to be present at any tip or internal node in the tree).

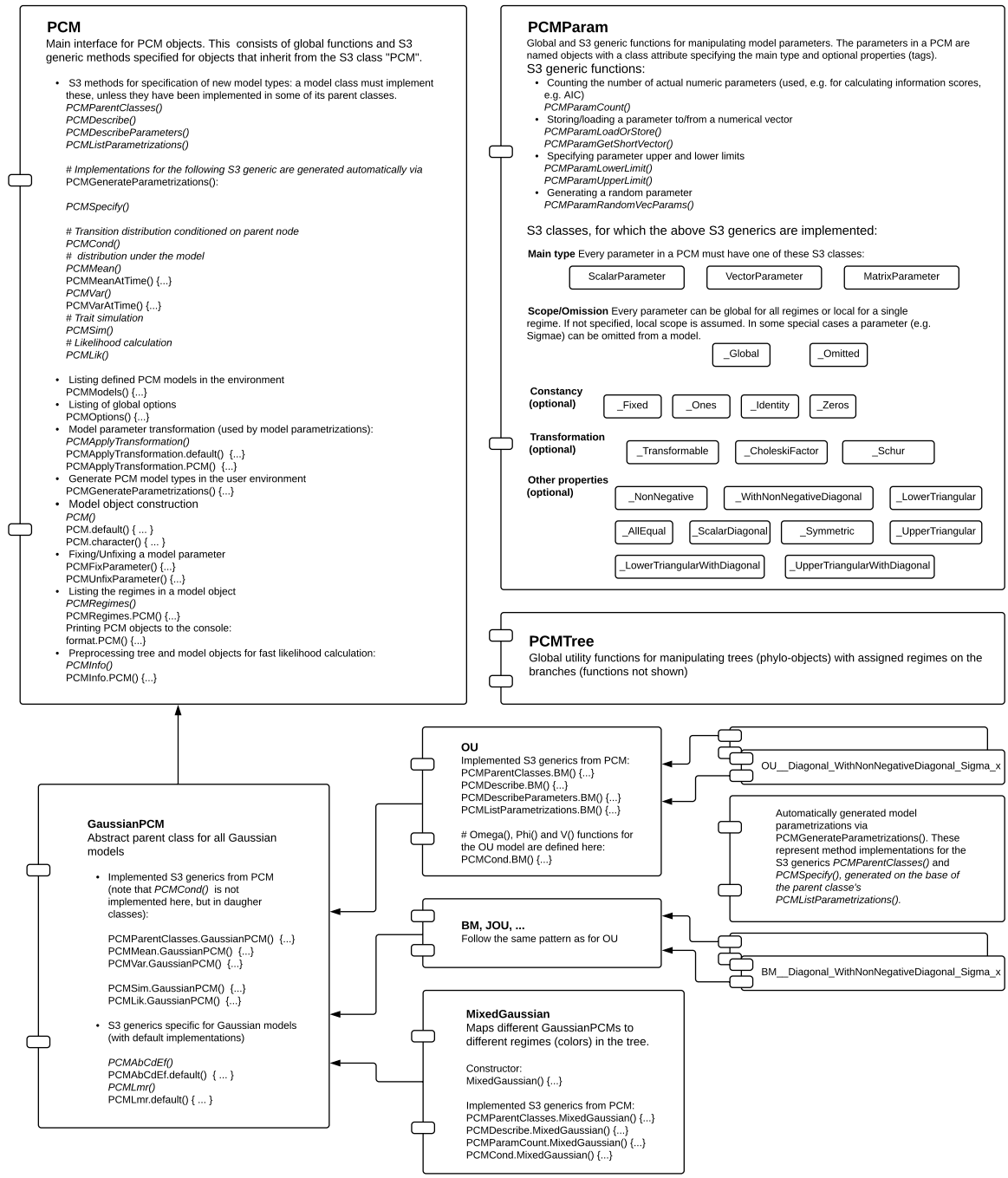
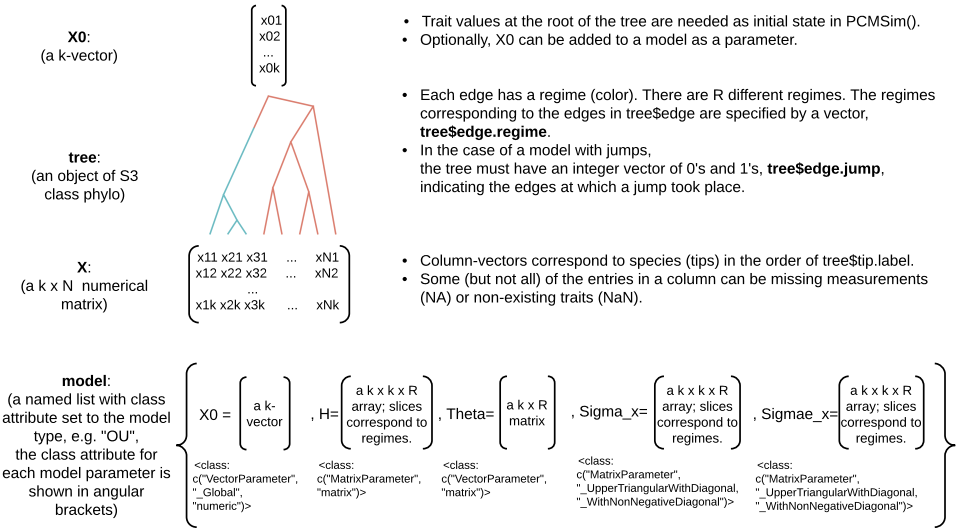


Figure 6.3: **An overview of the PCMBase package.** Each box represents a module. The modules "PCM", "PCMParam" and "PCMTree" define the end-user interface. In particular, the module "PCM" defines the interface for adding model extensions. Function names written in *italic* style denote S3 generic declarations. These functions can be defined or overwritten by inheriting modules, to provide model-specific behavior. The module "GaussianPCM" implements the pruning-wise likelihood evaluation. The functions $\vec{\omega}$, $\mathbf{\Phi}$ and $\mathbf{V}$ for each model within the framework must be implemented in specifications of the S3 generic function "PCMCond". It is possible to define parametrizations restricting particular model parameters, e.g. forcing a matrix parameter to be a diagonal matrix.



- **Showing the PCMBase runtime options:**

```
PCMOptions()
```

- **Showing the currently available model types**

```
PCMModels()
```

- **Creating a PCM object for 2 traits and 3 regimes**

```
# This will create a model object with default parameter values
model <- PCM("OU", k=2, regimes = 1:3)
print(model)
```

- **Setting parameter values manually**

```
# The brackets in model$X0[ ]  below are important
model$X0[ ] <- c(0.2, 1.3)
# regime 1:
model$H[,, 1] <- matrix(c(1, 0, 0, 1), 2, 2)
model$Sigma_x[,, 1] <- matrix(c(1, 0, 0.6, 1), 2, 2)
# regime 2:
model$H[,, 2] <- matrix(c(2, 0, 0, 1), 2, 2)
model$Sigma_x[,, 2] <- matrix(c(1, 0, 0.6, 1), 2, 2)
# regime 3:
model$H[,, 3] <- matrix(c(0.12 0, 0, 0.2), 2, 2)
model$Sigma_x[,, 3] <- matrix(c(1, 0, 0.6, 1), 2, 2)

print(model)

# CAUTION: Do not assign entire parameters directly, because
# this will eraze the parameter's attributes including its class.
# The following code is wrong:
model$X0 <- c(0.2, 1.3)
```

- **Setting random parameter values**

```
# First, we generate a random parameter vector for the model.
# The following code generates a random parameter vector for
# the model  drawing from a uniform distribution between the
# model's lower and upper limits:
vecParamValues <- PCMParamRandomVecParams(model)
print(vecParamValues)
# then, we load the generated vector into the model
PCMParamLoadOrStore(model, vecParamValues, 0, load=TRUE)
print(model)
```

- **Simulating a tree with regimes (using the phytools package)**

```
# make results reproducible
set.seed(20)

R <- 3; N <- 200;
# rate mtrix of transition from one regime to another
Q <- matrix(c(-0.1, 0, 0.1, 0, 0.01, -0.01, 0.1, 0, -0.1),
            R, R)
colnames(Q) <- rownames(Q) <- letters[1:R]
tree <- phytools::pbtree(n=N, scale=1, b = 1, d = 0.4)
tree.abc <- phytools::sim.history(tree, Q, anc='a')
tree.abc.s <- phytools::map.to.singleton(tree.abc)
tree.abc.s$edge.regime <- match(names(tree.abc.s$edge.length),
      letters[1:R])
PCMTreePlot(tree.abc.s, layout="fan")
```

- **Simulating trait evolution**

```
# The following code generates a k x M matrix, where M is the total
# number of nodes in the tree:
data <- PCMSim( tree.abc.s, model, X0=model$X0 )
# The first columns correspond to the tips in the order of tree$tip.label
X <- data[ , 1:PCMTreeNumTips(tree.abc.s)]
PCMPlotTraitData2D(X, tree.abc.s)
```

- **One likelihood calculation on a given model**

```
loglik <- PCMLik(X, tree.abc.s, model)
```

- **Multiple likelihood calculations on different model parameters**

```
# The first N columns correspond to the tips
X <- data[ , 1:PCMTreeNumTips(tree.abc.s)]
if( require(PCMBaseCpp) == FALSE ){
  # using default R-implementation (PCMLmr.default):
  metaInfo <- PCMInfo(X, tree.abc.s, modelOU)
} else {
    # using fast (parallel) C++ implementation (PCMBaseCpp):
  metaInfo <- PCMInfoCpp(X, tree.abc.s, modelOU)
}

# main loop (e.g. during model inference on given data)
while( NOT FINISHED ) {
  loglik <- PCMLik(X, tree.abc.s, modelOU, metaI = metaInfo)
  model <- propose.next.model.of.the.same.class(model)
}
```

Figure 6.4: **Using the 'PCMBase' package** The main runtime objects are depicted on the top of the figure, followed by coding examples for the specific use-cases.

2. Then, the coefficients $\mathbf{A}_i$, $\vec{b}_i$, $\mathbf{C}_i$, $\vec{d}_i$, $\mathbf{E}_i$ and f_i are calculated for each internal and tip node in the tree based on the values $\vec{\omega}$, $\mathbf{\Phi}$ and $\mathbf{V}$ calculated in the previous step. This calculation is done in the function '`PCMAbCdEf()`' within the module "GaussianPCM" which, again, is inherited by all model modules (see Fig. 6.3).
3. Next, the coefficients $\mathbf{L}_i$, $\vec{m}_i$, r_i are calculated based on the trait values at the tips, the values of $\mathbf{A}_i$, $\vec{b}_i$, $\mathbf{C}_i$, $\vec{d}_i$, $\mathbf{E}_i$ and f_i calculated in the previous step, and the recursive procedure described in Section 6.2.4, Eqs. (6.10), (6.11) and (6.12).
4. Finally, the (log–)likelihood value is calculated using the formula

$$\ell(\Theta) = pdf(\mathbf{X}|\vec{x}_0, \mathbb{T}, \mathbf{\Theta}) = \exp\left(\vec{x}_0^T \mathbf{L}_0 \vec{x}_0 + \vec{x}_0^T \vec{m}_0 + r_0\right), \tag{6.13}$$

where Θ denotes the set of model parameters and $\vec{x}_0$ is treated either as a parameter (specified as a member '`X0`' in the model object) or as the optimum point of the above equation given by:

$$\vec{x}_0 = -0.5 L_0^{-1} \vec{m}_0. \tag{6.14}$$

6.3.3 *Parallel likelihood calculation with the '**PCMBaseCpp**' add–in*

For faster likelihood calculation, it is possible to use multiple processor cores to perform the calculation of $\vec{\omega}$, $\mathbf{\Phi}$, $\mathbf{V}$, $\mathbf{A}_i$, $\vec{b}_i$, $\mathbf{C}_i$, $\vec{d}_i$, $\mathbf{E}_i$ and f_i in parallel. This is possible, given the fact that these coefficients depend solely on the model parameters and on the branch lengths in the tree (see, e.g. Eqs. (6.17) and (6.18)). The calculation of the coefficients $\mathbf{L}_i$, $\vec{m}_i$, r_i is not fully parallelizable but can be divided in parallelizable steps (generations) using a parallel post–order traversal algorithm (Mitov and Stadler, 2017a). We implemented this idea in the accompanying package '**PCMBaseCpp**', built on top of the '**Armadillo**' template library for linear algebra (Sanderson and Curtin, 2016), the '**Rcpp**' package for seamless '`R`' and '`C++`' integration (Eddelbuettel, 2013) and the '**SPLITT**' library for parallel tree traversal (Mitov and Stadler, 2017a).

We compared the performance of the multivariate serial and parallel '**PCMBase**' implementation against other univariate and multivariate implementations in a separate work (Mitov and Stadler, 2017a). As shown in (Mitov and Stadler, 2017a), on contemporary multi-core CPUs, the parallel '**PCMBaseCpp**' implementation can speed up the likelihood calculation up to an order of magnitude starting with 2 traits and trees of 100 to 10'000 tips. For univariate OU models, it can be beneficial to implement stand-alone classes bypassing the complex $k \times k$ matrix operations involved in the multivariate case. As shown in (Mitov and Stadler, 2017a), this can result in up to 100 fold faster likelihood calculation in the stand-alone class implementation. The use of PCMBaseCpp as a C++ back-end is recommended even if not using multi-core parallelization, because serial C++ code execution is still nearly 100 times faster than the equivalent implementation written in R (R-version at time of writing this article was 3.5).

6.4 STANDARD EXTENSIONS

6.4.1 *Missing values*

The trait measurement data are the observations at the tips. If a tip is described by a suite of traits it can easily happen that some of them are missing, either due to missing measurement

or because the corresponding trait does not exist for the species. Removing such a tip from any further analysis would be wasting information, i.e. the observed data for the tip. We notice that missing measurements for existing traits correspond to the marginal distribution of the observed measurements. In contrast, non-existing traits correspond to reduced dimensionality of the trait vector for the tip in question. Our computational framework keeps track of both of these cases by carefully accounting for the dimensionality of the trait vectors at the tips and the internal nodes and appropriately marginalizing during the integration part, as described below (see also Thm. 4 for examples). The input data is passed as a matrix (rows—trait measurements, columns—different species) the missing measurements have to be indicated as '`NA`'s, whereas the non-existing traits have to be indicated as '`NaN`'s (fig. 6.1).

We now turn to describing the technicalities of the mechanism taking care of the missing data. We use a vector of positive integers, $\vec{k}_j$, to denote the ordered set of active coordinates for a node j. If j is a tip, then $\vec{k}_j$ gives the indices of all non–missing entries in the trait vector for j; for an internal (unmeasured) node this gives the possibility to make some trait inactive. The cardinality of a vector is denoted with $|\vec{k}|$. For a vector, the notation $\vec{\theta}[\vec{k}]$ means the vector of elements of $\vec{\theta}$ on the coordinates contained in $\vec{k}$, while for a matrix $\mathbf{H}[\vec{k}_1, \vec{k}_2]$ means the matrix $\mathbf{H}$ with only the rows on the coordinates contained in $\vec{k}_1$ and columns contained in $\vec{k}_2$. For example take $\vec{\theta} = (10, 11, 12, 13)$ and $\vec{k} = (1, 3)$, then $\vec{\theta}[\vec{k}] = (10, 12)$, while if $\vec{k}_1 = (1, 3)$, $\vec{k}_2 = (2, 4)$ and

$$\mathbf{H} = \begin{bmatrix} 10 & 11 & 12 & 13 \\ 14 & 15 & 16 & 17 \\ 18 & 19 & 20 & 21 \\ 22 & 23 & 24 & 25 \end{bmatrix},$$

then

$$\mathbf{H}[\vec{k}_1, \vec{k}_2] = \begin{bmatrix} 11 & 13 \\ 19 & 21 \end{bmatrix}.$$

If a vector or matrix does not have any indication on which entries it is retained, then it means that we use the whole vector or matrix. All of the above notation is graphically represented in Fig. 6.1.

In Thm. 2 we showed that in our framework we have the representation that $\vec{x}_i \in \mathbb{R}^{k_i}$ conditional on $\vec{x}_j \in \mathbb{R}^{k_j}$ is $\mathcal{N}(\vec{\omega}_i + \mathbf{\Phi}_i \vec{x}_j, \mathbf{V}_i)$ distributed. Here, there is no issue on missing values as we are just working with a probabilistic representation. However, in practice one assumes some stochastic model for a k dimensional trait and under it (and that all observations are of full dimension) we would have that $\vec{x}_i \in \mathbb{R}^k$ conditional on $\vec{x}_j \in \mathbb{R}^k$ is $\mathcal{N}(\tilde{\vec{\omega}}_i + \tilde{\mathbf{\Phi}}_i \vec{x}_j, \tilde{\mathbf{V}}_i)$ distributed, for some auxiliary matrices $\tilde{\mathbf{\Phi}}_i$, $\tilde{\mathbf{V}}_i$ and vector $\tilde{\vec{\omega}}_i$. Then, to obtain the required representation of Thms. 2 and 3 we set

$$\begin{aligned} \vec{\omega}_i &= \tilde{\vec{\omega}}_i[\vec{k}_i], \\ \mathbf{\Phi}_i &= \tilde{\mathbf{\Phi}}_i[\vec{k}_i, \vec{k}_j], \\ \mathbf{V}_i &= \tilde{\mathbf{V}}_i[\vec{k}_i, \vec{k}_i]. \end{aligned} \tag{6.15}$$

6.4.2 *Measurement error*

Commonly in PCMs the observed values at the tips are averages from a number of individuals of each species. Using just these average values does not take into account the intra–species variability. Ignoring this can have profound effects on any further estimation (see Hansen and Bartoszek, 2012). Following the PCM tradition, we call this intra–species variability a measurement error, but one should remember that it can be due to true biological variability. Including this variability in our framework is straightforward. One recognizes, which component of the quadratic polynomial representation corresponds to the variance of the tip and augments it by the measurement error variance matrix, see the formulae in Section 6.5. From the user interface point of view this is a bit more complicated. The measurement error variance matrix is specific to each tip. Therefore in this situation the user has to define for each tip a different regime, with a regime specific variance matrix (called '`Sigmae`' in the implemented by us classes). Of course other model parameters can also be regime specific, e.g. the deterministic optima ('`Theta`' in the implemented by us classes).

6.4.3 *Non–ultrametric trees and multifurcations*

If one has only measurements from contemporary species, then the phylogeny describing them is naturally an ultrametric one. However, if for some reason the phylogeny is not ultrametric, e.g. it contains extinct species, then the quadratic polynomial framework can be directly employed. Because each branch is treated separately, it does not matter whether the tree is or is not ultrametric. Therefore, there is no need to search for transformations as in the 3–point structure based methods. This we believe should make the '**PCMBase**' package very straightforward to use. Furthermore, from the proof of Thm. 3 it is obvious that the tree does not need to be binary. Therefore, this adds even more flexibility to the user, they may use trees with polytomies.

6.4.4 *Punctuated equilibrium*

It is an ongoing debate in evolutionary biology whether the dominant mode of evolution is a gradual one or that during brief periods of time species undergo rapid change. Any gradual model of evolution can be extended to have a punctuated component by including jumps. Jump mechanisms, like jumps at the start of specific lineages or common jumps for daughter lineages, have to be developed on a per model basis, see Section 6.5.3 for an example. One current restriction is that '**PCMBase**' assumes that lineages do not interact after speciation. It is not possible to implement a model class such that if one daughter lineage jumps the other does not (this is communication between lineages after speciation). Therefore, to have such a situation the user needs to by themselves code on which lineages a jump can take place and on which it cannot. This can be easily achieved using the jumps mechanism of '**PCMBase**'. The '`phylo`' phylogenetic tree object can be enhanced by a '`edge.jump`' binary vector. The length of this vector equals the number of edges in the tree. A 0 entry indicates that no jump took place on the corresponding branch, while a 1 entry that it did.

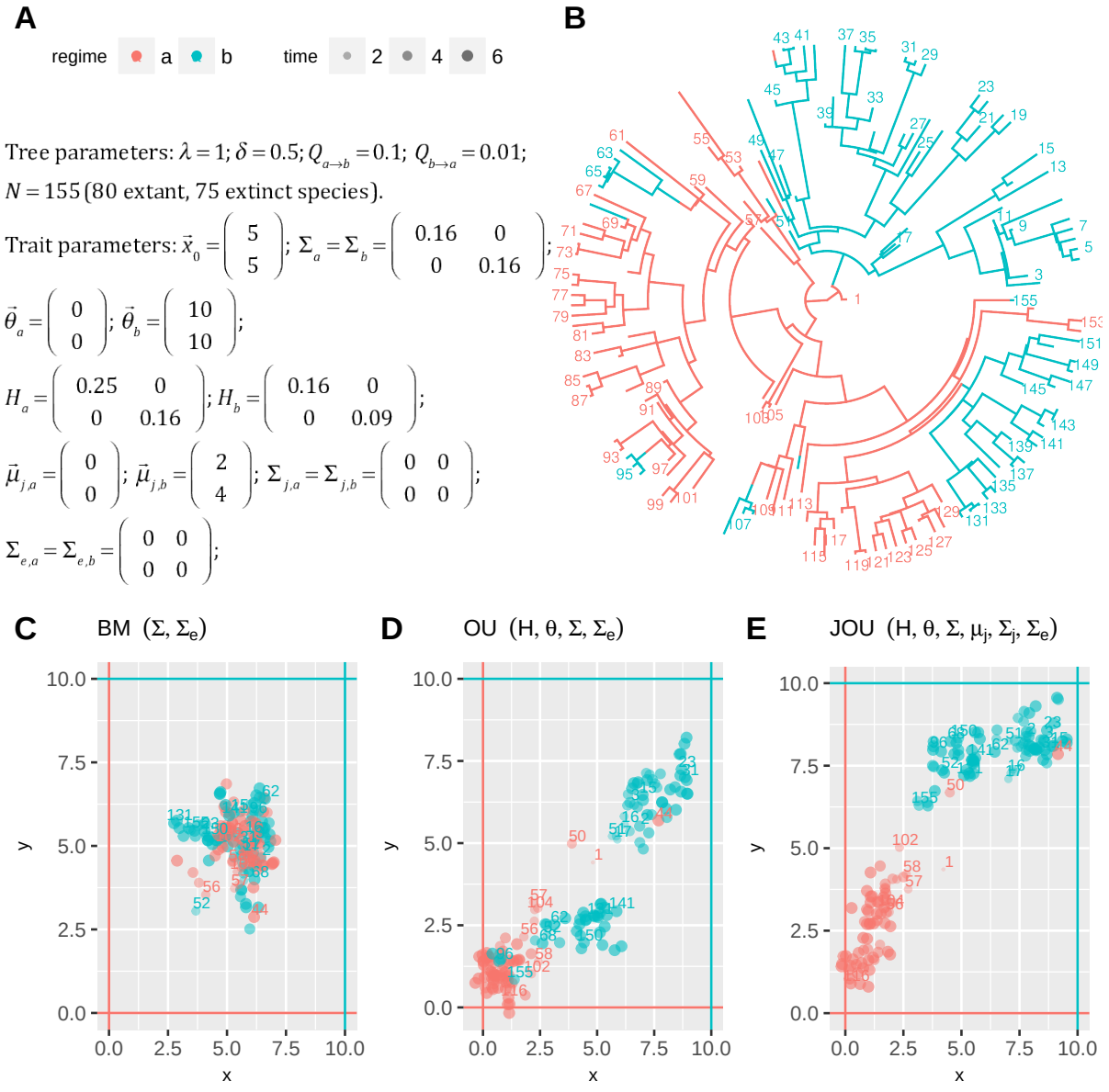


Figure 6.5: Simulations of trait evolution under four PCMs. A: parameters of the tree simulation (λ: speciation–rate; δ: extinction–rate; $Q_{a\to b}$: migration rate from habitat "a" to habitat "b"; $Q_{b\to a}$: vice–versa of $Q_{a\to b}$. The other parameters are described in the text. B: A birth–death phylogenetic tree generated using the function '`pbtree()`' and '`sim.history()`' from the package '**phytools**' (Revell, 2011). C—E: scatter plots of the traits observed at the tips of the tree after random simulation using the function '`PCMSim()`' of the '**PCMBase**' package.

6.5 ORNSTEIN–UHLENBECK TYPE MODELS

6.5.1 *The phylogenetic Ornstein–Uhlenbeck process*

Currently the Ornstein–Uhlenbeck process is the workhorse of the phylogenetic comparative methods framework. Since its introduction by Hansen (1997) it has been considered in detail with multiple software implementations (e.g. Bartoszek et al., 2012; Beaulieu et al., 2012; Butler and King, 2004; Clavel, Escarguel, and Merceron, 2015; FitzJohn, 2010; Goolsby, Bruggeman, and Ané, 2016; Hansen, Pienaar, and Orzack, 2008; Ho and Ané, 2014a, to name a few)

In the most general form, the multivariate Ornstein–Uhlenbeck process describes the evolution of a k–dimensional suite of traits $\vec{x} \in \mathbb{R}^k$ over a period of time by the following stochastic differential equation

$$\mathrm{d}\vec{x}(t) = -\mathbf{H}\left(\vec{x}(t) - \vec{\theta}(t)\right)\mathrm{d}t + \mathbf{\Sigma}_x \mathrm{d}\vec{W}(t), \tag{6.16}$$

where $\vec{W}(t)$ is a k–dimensional standard Wiener process, $\mathbf{H} \in \mathbb{R}^{k\times k}$, $\vec{\theta}(t) \in \mathbb{R}^k$ and $\mathbf{\Sigma}_x \in \mathbb{R}^{k\times k}$. Notice that when $\mathbf{H} = \mathbf{0}$, we obtain a Brownian motion model.

There is no current software package, in the case of phylogenetic OU models, that allows for an arbitrary form of the matrix $\mathbf{H}$. Except for the Brownian motion case, nearly all assume that $\mathbf{H}$ has to be symmetric–positive–definite (note that this encompasses the single trait case). '**mvMORPH**' (Clavel, Escarguel, and Merceron, 2015), '**SLOUCH**' (Hansen, Pienaar, and Orzack, 2008) and '**mvSLOUCH**' (Bartoszek et al., 2012) seem to be the only exceptions. '**mvMORPH**' and '**mvSLOUCH**' allow for a general invertible $\mathbf{H}$ (with options to restrict it to diagonal, triangular, symmetric positive–definite, positive eigenvalues, real eigenvalues or generally invertible). Furthermore, '**mvSLOUCH**' allows for a special singular structure of $\mathbf{H}$. The matrix has to have in the upper–left–hand corner an invertible matrix ('**SLOUCH**', the univariate predecessor of '**mvSLOUCH**' has a scalar here), arbitrary values to the right and $\mathbf{0}$ below. This type of model is called an Ornstein–Uhlenbeck–Brownian motion (OUBM) model. In contrast when $\mathbf{H}$ is non–singular the model is called an Ornstein–Uhlenbeck–Ornstein–Uhlenbeck (OUOU) one, some variables are labelled as predictors while the rest as responses.

It is of course not satisfactory to have restrictions on the form of $\mathbf{H}$. Different setups have different biological interpretations with regards to modelling causation (see Bartoszek et al., 2012; Reitan, Schweder, and Henderiks, 2012). In particular singular matrices will be interesting as they will correspond to certain linear combinations of traits under selection pressures while other linear combinations are free of this. The OUBM model is a special case where a pre–defined group of traits is assumed to evolve marginally as a Brownian motion. Of course a more general setup is desirable and actually, as we show in this work, possible.

Here the the only assumption we make on $\mathbf{H}$ is that it posses an eigendecomposition, $\mathbf{H} = \mathbf{P}\mathbf{\Lambda}\mathbf{P}^{-1}$ ($\mathbf{\Lambda}$ is a diagonal matrix, and the i–th element of the diagonal is denoted as λ_i). In particular $\mathbf{\Lambda}$ can be singular, i.e. some eigenvalues are 0 and furthermore the eigenvalues/eigenvectors are allowed to be complex.

In this work we assume that $\mathbf{\Sigma}_x$ is upper triangular (alternatively lower). Despite how it looks at first sight, this is not any sort of restriction, as in the likelihood we have only $\mathbf{\Sigma} := \mathbf{\Sigma}_x\mathbf{\Sigma}_x^T$. We furthermore assume that $\mathbf{\Sigma}$ is non–singular, otherwise the whole model would be singular from a statistics point of view.

Most OU model implementations assume that the deterministic optimum $\vec{\theta}_t$ is constant along each branch. Different branches may have different levels of it but regime switches along a branch are not allowed (however, Bastide et al., 2018; Ingram and Mahler, 2013; Khabbazian et al., 2016, are exceptions as they attempt to infer points of switching).

If we assume that the process starts at a value $\vec{x}(0) = \vec{x}_0$, then after evolution over time t (assuming all parameters are constant on this interval) it will be normally distributed with mean vector and variance–covariance matrix (Eqs. (A.1, B.2) Bartoszek et al., 2012)

$$\begin{aligned} \mathrm{E}\left[\vec{x}\right](t) &= e^{-\mathbf{H}t}\vec{x}_0 + \left(\mathbf{I} - e^{-\mathbf{H}t}\right)\vec{\theta} \in \mathbb{R}^k \\ \mathrm{Var}\left[\vec{x}\right](t) &= \int_0^t e^{-\mathbf{H}v}\boldsymbol{\Sigma}e^{-\mathbf{H}^T v}\mathrm{d}v \\ &= \mathbf{P}\left(\left[\frac{1}{\lambda_i+\lambda_j}\left(1-e^{-(\lambda_i+\lambda_j)t}\right)\right]_{1\leq i,j\leq k} \odot \mathbf{P}^{-1}\boldsymbol{\Sigma}\mathbf{P}^{-T}\right)\mathbf{P}^T \equiv \mathbf{V}(t) \in \mathbb{R}^{k\times k}, \end{aligned} \tag{6.17}$$

where $\mathbf{I}$ is the identity matrix of appropriate size. Notice that in the above, $\mathbf{H}$ only enters the moments, through its exponential. Therefore the moments can be calculated (and hence the distribution is well defined) for all $\mathbf{H}$, including defective ones. However, if $\mathbf{H}$ has (as we assumed) an eigendecomposition, then the exponential and in turn variance formula can be calculated effectively. If $\lambda_i = \lambda_j = 0$, then the term in the variance has to be treated in the limiting sense $\lambda^{-1}(1-e^{-\lambda t}) \to t$ with $\lambda \to 0$. Therefore, the variance matrix is always well defined and never singular for $t > 0$.

We assumed that $\mathbf{H}$ has to have an eigendecomposition while the process is well defined for any $\mathbf{H}$, including defective ones. Calculation of the matrix exponential for a defective matrix can be done using Jordan block decomposition. However, we do not provide such functionality, as Jordan block decomposition is numerically unstable and in fact, we are not aware of any 'R' implementation of it. Hence, defective matrices will result in errors. However, it is important to remember that defectiveness is the exception and not the rule for matrices. If checked for (by e.g. checking if the eigenvector matrix from '`eigen()`"s output is non–singular, Corollary 7.1.8., p. 353 Golub and Van Loan, 2013) and handled before calling using our package, it should not cause major issues. why don't we state what $V, \Psi, omega$ is?

6.5.2 *Multivariate Ornstein–Uhlenbeck*

The Ornstein–Uhlenbeck is a special case of Eq. (6.1) and here due to its importance in the phylogenetic comparative methods community we discuss it in detail. In particular we show how to construct the composite parameters found in the proof of Thm. 2 from the OU process representation of Eq. (6.16).

To simplify notation we denote the defined in Eq. (6.17) covariance matrix as $\tilde{\mathbf{V}}_i \equiv \mathbf{V}(t_i) + \delta_{i\in\{\mathbb{T}_0' s\ tips\}}\boldsymbol{\Sigma}_e^i$, where the Kronecker δ–symbol is defined as $\delta_{i\in\{\mathbb{T}_0' s\ tips\}} = 1$ if i is a tip of the tree and $\delta_{i\in\{\mathbb{T}_0' s\ tips\}} = 0$. The matrix $\boldsymbol{\Sigma}_e^i$ is the measurement error or intra–species variability variance matrix for tip species i.

Theorem 4. *Let $\vec{k}_i$ be the vector of coordinates on which $\vec{x}_i$ is observed, $\vec{k}_j$ be the vector of coordinates for $\vec{x}_j$ and $\vec{k}$ the full vector of coordinates. Using the parametrization found in the proof of Thm. 2 a multivariate Ornstein–Uhlenbeck process of evolution can be represented as*

$$\begin{aligned} \mathbf{V}_i &= \tilde{\mathbf{V}}_i[\vec{k}_i, \vec{k}_i] \in \mathbb{R}^{|\vec{k}_i|\times|\vec{k}_i|}, \\ \vec{\omega}_i &= \left(\mathbf{I}[\vec{k}_i, \vec{k}] - e^{-\mathbf{H}t_i}[\vec{k}_i, \vec{k}]\right) \vec{\theta}_i[\vec{k}] \in \mathbb{R}^{|\vec{k}_i|}, \\ \mathbf{\Phi}_i &= e^{-\mathbf{H}t_i}[\vec{k}_i, \vec{k}_j] \in \mathbb{R}^{|\vec{k}_i|\times|\vec{k}_j|}. \end{aligned} \tag{6.18}$$

Proof. In the multivariate OU case, Eq. (6.1) will be

$$pdf(\vec{x}_i|\vec{x}_j, t_i) = \mathcal{N}\left(e^{-\mathbf{H}t_i}\vec{x}_j + \left(\mathbf{I}[\vec{k}_i, \vec{k}] - e^{-\mathbf{H}t_i}[\vec{k}_i, \vec{k}]\right)\vec{\theta}_i[\vec{k}], \mathbf{V}_i[\vec{k}_i, \vec{k}_i]\right).$$

□

These formulae do not depend on whether the eigenvalues of $\mathbf{H}$ are positive, negative or 0. They will still be correct. The exponentiation of $\mathbf{H}$ will also not depend on this. Only with $\mathbf{V}_i$ will we need to take an appropriate limit as an eigenvalue is 0, see comments after Eq. (6.17).

Corollary 1. *For a multivariate Brownian motion, $\mathbf{H} = \mathbf{0}$ and $\tilde{\mathbf{V}}_i = t_i\mathbf{\Sigma} + \delta_{i\in\{\mathbb{T}_0's\ tips\}}\mathbf{\Sigma}_e^i$ process of evolution, hence using the parametrization found in the proof of Thm. 2 one can represent it as*

$$\begin{aligned} \mathbf{V}_i &= \tilde{\mathbf{V}}_i[\vec{k}_i, \vec{k}_i] \in \mathbb{R}^{|\vec{k}_i|\times|\vec{k}_i|}, \\ \vec{\omega}_i &= \vec{0}[\vec{k}_i] \in \mathbb{R}^{|\vec{k}_i|}, \\ \mathbf{\Phi}_i &= \mathbf{I}[\vec{k}_i, \vec{k}_j] \in \mathbb{R}^{|\vec{k}_i|\times|\vec{k}_j|}. \end{aligned} \tag{6.19}$$

In Fig. 6.5, panel C one can see an example collection of tip observations resulting from simulating of a bivariate trait following a BM process on top of a phylogeny and in panel D following an OU process.

6.5.3 *Multivariate Ornstein–Uhlenbeck with jumps*

It is an ongoing debate in evolutionary biology at what time does evolutionary change take place. Two theories state that change may take place either at times of speciation (punctuated equilibrium Eldredge and Gould, 1972; Gould and Eldredge, 1993) or gradually accumulate (phyletic gradualism, see references in Eldredge and Gould, 1972). There seems to be evidence for both types of evolution. For example, Bokma (2002) discusses that punctuated equilibrium is supported by fossil records (see Eldredge and Gould, 1972) but on the other hand Stebbins and Ayala (1981) also indicate experiments supporting phyletic gradualism.

Therefore, one would want processes that incorporate both types of evolution and allow for testing if either of them dominates. Ornstein–Uhlenbeck with jumps models are a framework where this is possible. Shortly, along a branch the traits follows an OU process. But then just after speciation a jump in the traits' values can take place. Whether such a jump takes place on a given, some or all daughter lineages is up to the specific implementation of the framework. From the perspective of the '**PCMBase**' package the location of the jumps has to be provided. It is in fact also possible in our implementation, to place jumps at arbitrary

points inside a branch. Models for jump locations are at a different level of PCM modelling, then what '**PCMBase**' handles.

Ornstein–Uhlenbeck processes with jumps capture a key idea behind the theory of punctuated equilibrium. At an internal node in the tree something happens that drives species apart and then "The further removed in time a species from the original speciation event that originated it, the more its genotype will have become stabilized and the more it is likely to resist change." (Mayr, 1982). Between branching events (and jumps) we can have stasis—"fluctuations of little or no accumulated consequence" taking place (Gould and Eldredge, 1993). This corresponds well to an OU with jumps model. If the speed of convergence of the process is large enough, then the stationary distribution is approached rapidly and the stationary oscillations around the (constant) mean can be interpreted as stasis between jumps.

Corollary 2. *For a multivariate OU defined with jumps, jump distribution $\mathcal{N}(\vec{\mu}_J, \mathbf{\Sigma}_J)$ and denoting by the indicator ξ_i (we assume that the jumps are known) if a jump took place at the start of the branch leading to node i, we have*

$$\tilde{\mathbf{V}}_i = \int_0^{t_i} e^{-\mathbf{H}v}\mathbf{\Sigma}e^{-\mathbf{H}^T v}\mathrm{d}v + \xi_i e^{-\mathbf{H}t_i}\mathbf{\Sigma}_J e^{-\mathbf{H}^T t_i} + \delta_{i\in\{\mathbb{T}_0' s\ tips\}}\mathbf{\Sigma}_e^i. \tag{6.20}$$

Using the parametrization found in the proof of Thm. 2 one can represent it as

$$\begin{aligned}
\mathbf{V}_i &= \tilde{\mathbf{V}}_i[\vec{k}_i, \vec{k}_i] \in \mathbb{R}^{|\vec{k}_i|\times|\vec{k}_i|},\\
\vec{\omega}_i &= \xi_i e^{-\mathbf{H}t_i}[\vec{k}_i, \vec{k}]\vec{\mu}_J[\vec{k}] + \left(\mathbf{I}[\vec{k}_i, \vec{k}] - e^{-\mathbf{H}t_i}[\vec{k}_i, \vec{k}]\right)\vec{\theta}_i[\vec{k}] \in \mathbb{R}^{|\vec{k}_i|},\\
\mathbf{\Phi}_i &= e^{-\mathbf{H}t_i}[\vec{k}_i, \vec{k}_j] \in \mathbb{R}^{|\vec{k}_i|\times|\vec{k}_j|}.
\end{aligned} \tag{6.21}$$

The multivariate Brownian motion with jumps model follows as an immediate corollary $(\mathbf{H} \to \mathbf{0})$.

Corollary 3. *For a multivariate Brownian motion with jumps (jumps defined the same as in Corollary 2) the variance at a node i is $\tilde{\mathbf{V}}_i = t_i\mathbf{\Sigma}[\vec{k}_i, \vec{k}_i] + \xi_i\mathbf{\Sigma}_J[\vec{k}_i, \vec{k}_i] + \delta_{i\in\{\mathbb{T}_0' s\ tips\}}\mathbf{\Sigma}_e^i$. Using the parametrization found in the proof of Thm. 2 one can represent it as*

$$\begin{aligned}
\mathbf{V}_i &= \tilde{\mathbf{V}}_i[\vec{k}_i, \vec{k}_i] \in \mathbb{R}^{|\vec{k}_i|\times|\vec{k}_i|},\\
\vec{\omega}_i &= \xi_i\vec{\mu}_J[\vec{k}_i] \in \mathbb{R}^{|\vec{k}_i|},\\
\mathbf{\Phi}_i &= \mathbf{I}[\vec{k}_i, \vec{k}_j] \in \mathbb{R}^{|\vec{k}_i|\times|\vec{k}_j|}.
\end{aligned} \tag{6.22}$$

In Fig. 6.5, panel E one can see an example collection of tip observations resulting from simulating of a bivariate trait following an OU process with jumps on top of a phylogeny.

6.5.4 *Beyond the Ornstein–Uhlenbeck process*

There are a number of popular PCM models that do not fall into the above described OU framework despite appearing very similar. In particular we mean the BM with trend, drift, early burst/Accelerating–decelerating (EB/ACDC) or white noise (implemented in the

'**geiger**' 'R' package Harmon et al., 2008). With the exception of white noise, they all can be represented by the SDE (cf. Eq. (1) of Manceau, Lambert, and Morlon, 2016)

$$\begin{cases} \mathrm{d}\vec{x}(t) &= \left(\vec{h}(t) - \mathbf{H}\vec{x}(t)\right)\mathrm{d}t + \mathbf{\Gamma}(t)\mathrm{d}\vec{W}(t), \\ \vec{x}(0) &= \vec{x}_0. \end{cases} \tag{6.23}$$

(Manceau, Lambert, and Morlon, 2016), then provide the expectation and variance under the model, by slightly modifying their Eqs. (4a) and (4b),

$$\begin{aligned} \mathrm{E}\left[\vec{x}_i|\vec{x}_j\right] &= e^{-t_i\mathbf{H}_i}\vec{x}_j + \int_{t_i^s}^{t_i^e} e^{(s-t_i^e)\mathbf{H}_i}\vec{h}_i(s)\mathrm{d}s, \\ \mathrm{Var}\left[\vec{x}_i|\vec{x}_j\right] &= \int_{t_i^s}^{t_i^e} e^{(s-t_i^e)\mathbf{H}_i}\mathbf{\Gamma}_i(s)\mathbf{\Gamma}_i^T(s)e^{(s-t_i^e)\mathbf{H}_i^T}\mathrm{d}s, \end{aligned} \tag{6.24}$$

where t_i^s is the time at the start of the branch and t_i^e at the end (of course $t_i = t_i^e - t_i^s$). This corresponds in our notation to

$$\begin{aligned} \vec{\omega}_i &= \int_{t_i^s}^{t_i^e} e^{(s-t_i^e)\mathbf{H}_i}\vec{h}_i(s)\mathrm{d}s, \\ \mathbf{\Phi}_i &= e^{-t_i\mathbf{H}_i}, \\ \mathbf{V}_i &= \int_{t_i^s}^{t_i^e} e^{(s-t_i^e)\mathbf{H}_i}\mathbf{\Gamma}_i(s)\mathbf{\Gamma}_i^T(s)e^{(s-t_i^e)\mathbf{H}_i^T}\mathrm{d}s. \end{aligned} \tag{6.25}$$

Naturally everything should be appropriately (as described in Section 6.4.1) adjusted if missing values are present. Hence, *in the subcase of non–interacting lineages*, our framework covers Manceau, Lambert, and Morlon (2016)'s. As the initially mentioned models are subcases (cf. Tab. 1 of Manceau, Lambert, and Morlon, 2016) they are available in our framework. In particular (after an appropriate generalization to the multivariate traits), ACDC model—$\vec{\omega}_i = \vec{0}$, $\mathbf{\Phi}_i = \mathbf{I}$, $\mathbf{V}_i = \int_{t_i^s}^{t_i^e} e^{s\mathbf{R}_i}\mathbf{\Sigma}_i\mathbf{\Sigma}_i^T e^{s\mathbf{R}_i^T}\mathrm{d}s$ (see Eq. 6.17 for how to calculate this integral when $\mathbf{R}$ is eigendecomposable), BM with drift—$\vec{\omega}_i = \vec{h}_i t$, $\mathbf{\Phi}_i = \mathbf{I}$, $\mathbf{V}_i = \mathbf{\Sigma}_i\mathbf{\Sigma}_i^T t_i$ and BM with trend—$\vec{\omega}_i = \vec{0}$, $\mathbf{\Phi}_i = \mathbf{I}$, $\mathbf{V}_i = \int_{t_i^s}^{t_i^e} \mathbf{\Gamma}_i(s)\mathbf{\Gamma}_i^T(s)\mathrm{d}s$, for a linear $\Gamma_i(s)$ (based on '**geiger**''s manual). The white noise process corresponds to a situation, where the observations are i.i.d.—a star phylogeny with all branches of length 1 and all species have same mean vector (denoted $\vec{h}$) and variance–covariance matrix ($\mathbf{V}$). In our representation this is $\vec{\omega}_i = \vec{h}$, $\mathbf{\Phi}_i = \mathbf{0}$ and $\mathbf{V}_i = \mathbf{V}$.

6.6 TECHNICAL CORRECTNESS

Validating the technical correctness is an important but often neglected step in the development of likelihood calculation software. This step is particularly relevant for complex multivariate models, because logical errors can occur in many levels, such as the mathematical equations for the different terms involved in the likelihood, the programming code implementing these equations, the code responsible for the tree traversal, the parametrization of the model and the preprocessing of the input data. These logical errors add up to numerical

errors caused by limited floating point precision, which can be extremely hard to identify. Ultimately, these errors lead to wrong likelihood values, false parameter inference and wrong analysis. All these concerns motivate for a systematic approach of testing the correctness of the software.

We implemented a technical correctness test of the three models currently implemented in '**PCMBase**' using the method of posterior quantiles proposed by Cook, Gelman, and Rubin (2006). The posterior quantiles method (Alg. 6.1) is a simulation based approach. It employs the fact that, for a fixed prior distribution of the model parameters, the sample of posterior quantiles of any model parameter, θ is uniform (see e.g. Cook, Gelman, and Rubin, 2006; Mitov and Stadler, 2017a, for details). Thus, any deviation from uniformity of the posterior quantile sample for any of the model parameters indicates the presence of an error, either in the simulation software, or in the likelihood calculator used to generate the posterior samples.

Algorithm 6.1 : Posterior quantiles method

1: Sample "true" parameters Θ from the prior;
2: Simulate random data, $\mathbf{X}_\Theta$, under the model specified by Θ;
3: Generate a sample S_θ from the posterior distribution $P_\theta = P(\theta|\mathbf{X}_\Theta)$;
4: Calculate the empirical quantile of the "true" θ in S_θ;

We used a fixed non–ultrametric tree of $N = 515$ tips with two regimes "a" and "b". The tree was generated using the functions '`pbtree()`' and '`sim.history()`' from the package '**phytools**' (Revell, 2011). We implemented the posterior quantile test using the '**BayesValidate**' 'R'–package (Cook, Gelman, and Rubin, 2006). For each model we set a parametrization and a prior distribution as follows:

- BM
  3 parameters: $\Theta_{BM} = [\Sigma_{11}, \Sigma_{12}, \Sigma_{e,11}]$, such that

$$\mathbf{\Sigma}_a = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12} & \Sigma_{11} \end{pmatrix}, \mathbf{\Sigma}_b = \begin{pmatrix} \Sigma_{11} & 0 \\ 0 & \Sigma_{11} \end{pmatrix}, \mathbf{\Sigma}_{e,a} = \mathbf{\Sigma}_{e,b} = \begin{pmatrix} \Sigma_{e,11} & 0 \\ 0 & \Sigma_{e,11} \end{pmatrix}$$

  prior: $\Sigma_{11} \sim \text{Exp}(1)$, $\Sigma_{12} \sim \mathcal{U}(-0.9\Sigma_{11}, 0.9\Sigma_{11})$, $\Sigma_{e,11} \sim \text{Exp}(10)$.

- OU
  8 parameters: $\Theta_{OU} = [\Theta_{BM}, \theta_{b,1}, \theta_{b,2}, H_{b,11}, H_{b,12}, H_{b,22}]$, such that $\mathbf{\Sigma}_a$, $\mathbf{\Sigma}_b$, $\mathbf{\Sigma}_{e,a}$ and $\mathbf{\Sigma}_{e,b}$ are defined as for BM and

$$\vec{\theta}_a = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \vec{\theta}_b = \begin{pmatrix} \theta_{b,1} \\ \theta_{b,2} \end{pmatrix}, \mathbf{H}_a = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}, \mathbf{H}_b = \begin{pmatrix} H_{b,11} & H_{b,12} \\ H_{b,12} & H_{b,11} \end{pmatrix}$$

  prior: for parameters in Θ_{BM} the same prior has been used as for the BM model; for the new parameters, the prior has been set as $\theta_{b,1} \sim \mathcal{N}(1, .25)$, $\theta_{b,2} \sim \mathcal{N}(2, .5)$, $H_{b,11} \sim \text{Exp}(1)$, $H_{b,22} \sim \text{Exp}(1)$, $H_{b,12} \sim \mathcal{U}(-0.9\sqrt{H_{b,11}H_{b,22}}, 0.9\sqrt{H_{b,11}H_{b,22}})$.

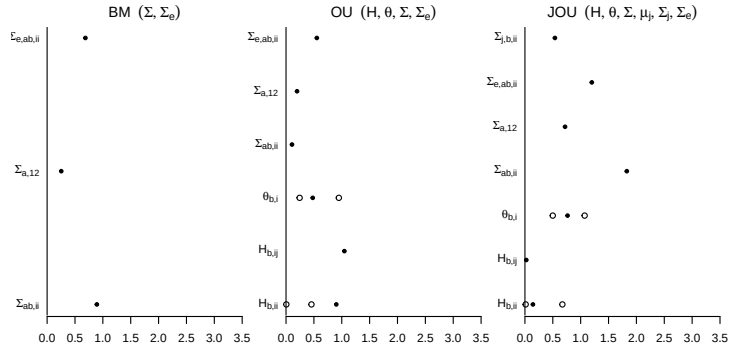


Figure 6.6: Absolute Z_θ–statistics for the four posterior quantile tests. The Z_θ statistic is described by Cook, Gelman, and Rubin (2006). High values indicate deviation from uniformity of the posterior quantile distribution for an individual model parameter (circles) or a batch of several model parameters (bullets). The reported values, smaller than 3 for all parameters, had insignificant p-values as well as Bonferroni–adjusted p–values. The plots were generated using the package '**BayesValidate**' (Cook, Gelman, and Rubin, 2006).

- JOU
  9 parameters: $\Theta_{JOU} = [\Theta_{OU}, \Sigma_{j,11}]$, such that $\mathbf{\Sigma}_a$, $\mathbf{\Sigma}_b$, $\mathbf{\Sigma}_{e,a}$, $\mathbf{\Sigma}_{e,b}$, $\vec{\theta}_a$, $\vec{\theta}_b$, $\mathbf{H}_a$, $\mathbf{H}_b$ are defined as for OU and

$$\vec{\mu}_{j,a} = \begin{pmatrix} -\theta_{b,1} \\ -\theta_{b,2} \end{pmatrix}, \vec{\mu}_{j,b} = \begin{pmatrix} \theta_{b,1} \\ \theta_{b,2} \end{pmatrix}, \mathbf{\Sigma}_{j,a} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \mathbf{\Sigma}_{j,b} = \begin{pmatrix} \Sigma_{j,11} & 0 \\ 0 & \Sigma_{j,11} \end{pmatrix}$$

  prior: for parameters in Θ_{OU} the same prior has been used as for the OU model; for the new parameter, the prior has been set as $\Sigma_{j,11} \sim \text{Exp}(10)$.

For each, model, we ran the function '`validate()`' from the '**BayesValidate**' package, setting the number of replications to 48. The results are summarized in Fig. 6.6. All Bonferroni adjusted p–values of the absolute Z_θ statistics were above 0.2, showing that the posterior quantiles did not deviate from uniformity (see Cook, Gelman, and Rubin, 2006, for details on Z_θ statistic).

6.7 DISCUSSION

Currently the mathematical frameworks proposed for PCMs are applied to situations that are very different from the original motivation of a between species analyses within a small clade of some quantitative trait. They are employed in many situations with a tree structure behind the measurements. For example, traits being gene expression levels (Bedford and Hartl, 2009; Rohlfs, Harrigan, and Nielsen, 2013) or epidemiological measurements (the tree connects the epidemic's outbursts Pybus et al., 2012) are analysed.

With large and diverse clades, there is a need to vary the parameters of the models across different clades or epochs in the tree. Already e.g. Bartoszek et al. (2012), Butler and King (2004), and Hansen (1997) showed the possibility of varying the deterministic optimum or OU processes. Beaulieu et al. (2012), Eastman et al. (2011), and Manceau, Lambert, and Morlon (2016) went further to allow all parameters of the underlying SDE to vary over the tree. Estimating the time and branches for parameter changes has been proposed in Ingram and Mahler (2013), Khabbazian et al. (2016) and Bastide et al. (2018) with implementations in '**SURFACE**', '*l***1ou**' and '**PhylogeneticEM**' 'R' respectively. The branching/extinction proba-

bilities may depend on the underlying trait (QuaSSE in '**diversitree**' FitzJohn, 2010). A different direction is the study of theoretical properties of branching stochastic processes. One can ask questions about the long term properties of the phylogenetic sample, parameter estimability or more generally of the distribution of the estimators (e.g. Ané, Ho, and Roch, 2016; Bartoszek and Sagitov, 2015; Cressler, Butler, and King, 2015; Ho and Ané, 2013, 2014b).

The situations mentioned above can easily require likelihood evaluations well beyond the amount of a "standard" optimization (e.g. an exponential number of regime patterns on the tree or reestimation to obtain the estimators' distribution). Furthermore, the trees connected to such calculations to be analysed can be huge, going into thousands of tips such as HIV data analysed e.g. in (Hodcroft et al., 2014; Mitov and Stadler, 2018). Hence, being able to quickly evaluate the likelihood is crucial.

Our package offers the possibility to very quickly obtain the likelihood for very large phylogenies for a wide range of models. Further, it is extremely flexible allowing the user to easily use it as a computational engine for their particular modelling setup/parametrization. '**PCMBase**' is able to handle multiple standard extensions, allowing the scientist to use all observed data. Finally, the package is written in such a way that it can be further developed to include more complex situations.

Some "standard extensions" from Section 6.4 deserve special mention. Firstly and briefly we remind the reader that as '**PCMBase**' handles non–ultrametric (Section 6.4.3) trees. Thus it can directly use fossil data or pathogen data. Currently, it is assumed that all samples are tips in the tree, thus we do not support sampled ancestor trees Gavryushkina et al., 2014. However, it will be straightforward to implement internal measurements on the tree if required by the user.

In the same Section 6.4.3, we notice that from the perspective of '**PCMBase**' the out–degree of an internal node is irrelevant. This is as the likelihood is calculated as the product over all daughter clades. Therefore, our computational engine should be appreciated by users who have poorly resolved trees with polytomies.

'**PCMBase**' handles incomplete observations of traits, meaning partially measured fossils do not pose any problem. As mentioned in Section 6.4.1 '**PCMBase**' distinguishes two types of missingness, unobserved trait ('`NA`') and non–existing trait ('`NaN`'). From the perspective of the user this might seem like a mere formality. However, from the perspective of the likelihood calculations it makes a profound difference. Unobserved traits are integrated over, meaning that first $\vec{\omega}_i$, $\mathbf{V}_i$, $\mathbf{\Phi}_i$ are calculated as if all k traits were present and only afterwards are appropriate entries/rows and columns removed. The second case of non–existing traits is treated differently, $\vec{\omega}_i$, $\mathbf{V}_i$, $\mathbf{\Phi}_i$ are calculated taking into account that the trait vector at the given node is from a lower dimension (i.e. $\mathbf{A}_i$, $\vec{b}_i$, $\mathbf{C}_i$, $\vec{d}_i$ and $\mathbf{E}_i$ are taken from lower dimensions by removing appropriate entries/rows and columns).

What should be particularly useful for the applied researcher is that one can specify the non–presence at internal nodes of the phylogeny. One does not need to have any measurements on these nodes. For example if one is studying five traits, one could have associated with an internal node j, the vector $\vec{k}_j = (\text{'}\texttt{NA,NA,NA,NaN,NaN}\text{'})^T$, meaning that there are no measurements on the first three traits, while the last two are not present at the species corresponding to node j. Firstly, this allows for correct handling of ancestral species that did not have exhibit certain traits present that are present in (some) contemporary species.

The mathematical approach utilized in the package is furthermore, very flexible in the sense that it can be directly extended in directions beyond the normal, non–interacting lineages setup.

A random variable with density as in Eq. (6.1) belongs to the quadratic exponential family as defined in Def. 2.

Definition 2. *(cf. Def. 2 of Gouri 'eroux, Monfort, and Trognon, 1984) A random variable is said to belong to the quadratic exponential family if its pdf is a function acting on* $\mathbb{R}^k$ *with representation*

$$pdf(\vec{u}) = \exp\left(a + b(\vec{u}) + \vec{c}\vec{u} + \vec{u}^T\mathbf{D}\vec{u}\right), \tag{6.26}$$

where a and $b(\vec{u})$ *are scalars,* $\mathbb{R}^k \ni \vec{c}$ *is a vector of size k and* $\mathbb{R}^{k\times k} \in \mathbf{D}$ *is a matrix.*

One can see that setting $b(\vec{u}) = 0$ (cf. Example 2.5 Ziegler, 2011) results in the representation of Eq. (6.1). However, the family of densities with pdfs following Eq. (6.26) is more general. We can see that from the proof of Thm. 3 that if in Thm. 2 we would drop the requirement of non–zero support on $\mathbb{R}^{k_i}$, then we can obtain non–normal models for whom the likelihood can be rapidly found using our method. If fact, we can see that the key step in Thm. 3's proof is the starred equality $\overset{\bigstar}{=}$. There we calculate the integral over $\mathbb{R}^{k_i}$. If the space is different, then in the box in the next step we would have a different constant. The only condition is that this constant (i.e. region of integration) cannot depend on $\vec{x}_j$. It is important to point out that our approach makes it straightforward to model traits that are constrained by some minimum and maximum value. We just include this assumption in the region of integration and constant f_i.

One can also include discrete models. The simplest example is a binary trait with states 0 and 1. Let $p_{ji}(t_i)$, $i \in \{0,1\}$, $j \in \{0,1\}$ be the probability of change from state j to i in time t_i. We can write this model in the form of Eq. (6.1),

$$\begin{aligned} pdf(x_i|x_j,t_i) = \exp\big(&\log p_{00}(t_i) + x_i(\log p_{01}(t_i) - \log p_{00}(t_i)) + x_j(\log p_{10}(t_i) - \log p_{00}(t_i)) \\ &+ x_i x_j(\log p_{00}(t_i) + \log p_{11}(t_i) - \log p_{01}(t_i) - \log p_{10}(t_i))\big), \end{aligned}$$

where $x_i, x_j \in \{0,1\}$. If furthermore, $p_{01}(t_i) = 1 - p_{11}(t_i)$, then the model will allow for a version of Thm. 3, i.e. the likelihood for all the observations will be of the form of Eq. (6.7). This is a big restriction as it implies that the transition matrix is parametrized by one number and has to be of the form

$$\begin{bmatrix} p_{00}(t) & 1 - p_{00}(t) \\ p_{00}(t) & 1 - p_{00}(t) \end{bmatrix}.$$

However, this is the price to pay so that the constant after the equality $\overset{\bigstar}{=}$ does not depend on x_j.

What we presented above with the binary trait is more of an mathematical exercise, to present an example of a non–normal model. Of course, Felsenstein (1973)'s pruning algorithm is the way to handle discrete models. There one needs to calculate the transmission probability $P(t)$ over a branch of length t. We assume a very special form for $P(t)$, allowing for rapid computations, and hence can only handle a very restricted sub–model.

Another direction in which '**PCMBase**' has the potential to be extended is to drop the assumption of independent lineage evolution after speciation, i.e. the trait exactly follows the tree structure. Such a restriction is of course not biologically realistic, but had to made in nearly all PCM inference packages due to complications caused for likelihood evaluation. In many species, especially plants, hybridization events take place. Furthermore, if the

tips of the tree correspond not to species but to populations, gene exchange can be continuous in time, i.e. migration takes place. Models with interactions, especially migration have been considered in the literature (e.g. Bartoszek et al., 2017; Drury et al., 2016; Jhwueng and OMeara, 2015; Manceau, Lambert, and Morlon, 2016; Nuismer and Harmon, 2015). From the mathematical point of view it is possible to include migration models in the described here framework. If the transition law is normal, one would consider the collection of co–evolving species and integrate over their joint ancestral state(s) (similarly as Bartoszek et al., 2017; Drury et al., 2016; Manceau, Lambert, and Morlon, 2016, treat the mean vector and variance–covariance matrix). However, from the implementation point of view it is more complex as one needs to keep track of which parts of the phylogeny are lumped and different model parameters for the lumped parts, that also describe how the lineages interact. Also, a user–friendly interface is a challenge. Therefore, we leave modelling interacting lineages for future developments of '**PCMBase**'.

Despite the generality, speed and easiness of use of the package the user has to be aware of a potential pitfall. Theorem 2 and the proof of Thm. 3 indicate a numerical weakness of our method. If a branch ending at node i is extremely short, then the associated with it variance–covariance matrix, $\mathbf{V}_i$, can be computationally singular. Hence, calculating its inverse, a necessary step to obtain the likelihood, will not be possible. '**PCMBase**' catches such an error and returns it, pointing to the offending node. '**PCMBase**' proposes a way to handle this condition: if the branch is shorter than a user-specified threshold (runtime options "PCMBase.Skip.Singular" and "PCMBase.Threshold.Skip.Singular"), the whole branch can be treated as a 0-length branch and skipped during the likelihood calculation.

ACKNOWLEDGMENTS

KB was supported by the Knut and Alice Wallenbergs Foundation, the G S Magnuson Foundation of the Royal Swedish Academy of Sciences (grant no. MG2016–0010) and is supported by the Swedish Research Council (Vetenskapsrådet) grant no. 2017–04951.

7

MIXED GAUSSIAN PHYLOGENETIC MODELS

Manuscript submitted for peer review as

Venelin Mitov, Krzysztof Bartoszek and Tanja Stadler (2018). Automatic Generation of Evolutionary Hypotheses using Mixed Gaussian Phylogenetic Models *Proceedings of the National Academy of Sciences (PNAS)*.

This chapter introduces mixed Gaussian phylogenetic models (MGPMs), based on the $\mathcal{G}_{LInv}$ family discussed in Chapter 6. MGPMs address one issue that has been stated in the discussion of Chapter 3, namely, the fact that most of the models used to estimate phylogenetic heritability assumed a homogeneous evolutionary process over the whole tree. Taking advantage of the PCMBase R-package for fast likelihood calculation, I develop an algorithm for approximate maximum likelihood inference of an optimal MGPM model fit to multiple trait phylogenetically linked comparative data. I illustrate this approach with an analysis of the brain–body–mass allometry in mammals.

ABSTRACT

Gaussian phylogenetic models like Brownian motion and Ornstein-Uhlenbeck processes are the workhorses of modeling continuous trait evolution. However, these models fit poorly to big trees, because they neglect the heterogeneity of the evolutionary process in different lineages of the tree. Previous works have addressed this issue by introducing shifts in the evolutionary model occurring at inferred points in the tree. In all current implementations, though, these shifts are "intra-model", meaning that they allow a jump in one or two model parameters, keeping all other parameters "global" for the entire tree. Such restrictions are artificial, because they are driven by non-biological concerns, e.g. computational feasibility. There is no biological reason to restrict a shift to a single model parameter or, even, to a single type of model. Mixed Gaussian phylogenetic models (MGPMs) incorporate the idea of jointly inferring different types of Gaussian models, with independent parameter sets, modeling the evolution in different parts of the tree. Here, we propose a new approximate maximum likelihood method for fitting MGPMs to comparative data comprising possibly incomplete measurements for several traits from extant and extinct phylogenetically linked species. The method enables data-driven generation of evolutionary hypotheses, reducing the need of uninformed preliminary modeling assumptions. We applied the method to the largest published tree of mammal species with body- and brain-mass measurements, showing strong statistical support for a MGPM with twelve distinct evolutionary regimes. Based on this result, we state a hypothesis for the evolution of the brain-body-mass allometry over the past 160 million years.

7.1 INTRODUCTION

Life is extremely diverse as the result of the dynamic change in evolutionary forces driving speciation and phenotypic evolution Benton and Emerson, 2007. Gaussian phylogenetic models, such as Brownian motion (BM) and Ornstein-Uhlenbeck (OU) processes, have become a standard tool in the comparative analysis of quantitative traits (Butler and King, 2004; Pennell and Harmon, 2013). Among many applications, these models have been used for correcting the errors from phylogenetic correlation in comparative regression analysis (Felsenstein, 1985; Martins and Hansen, 1997), for quantifying the phylogenetic signal in morphological and pathogen traits (Alizon et al., 2010; Bertels et al., 2017; Blanquart et al., 2017; Hodcroft et al., 2014; Housworth, Martins, and Lynch, 2004; Mitov and Stadler, 2018; Shirreff et al., 2013), and for testing hypotheses about the evolutionary forces that have seeded the patterns in the traits observable nowadays (Butler and King, 2004; Hansen and Martins, 1996; Pennell and Harmon, 2013).

With ever growing tree size and scope of the phylogenetic analysis, it is unlikely that a single regime of evolution described by a single model could have driven the changes in the traits across the entire tree. Such a model would have too low of a resolution to accommodate the inherent heterogeneity in the evolutionary process. Even worse, fitting a misspecified model to a large phylogeny is prone to inferring statistically significant, but strongly biased parameter values, due to their tendency to "compensate" for the modeling error (Cooper et al., 2015; Mitov and Stadler, 2018). There is no biological reason to constrain the change of a model regime to a single or a few model parameters, nor is there any reason to restrict the change to a single type of model. However, to the best of our knowledge, all current implementations inferring phylogenetic models with shifts, impose such restrictions, motivated mainly by computability issues (Bastide, Mariadassou, and Robin, 2017; Bastide

et al., 2018; Beaulieu et al., 2012; Eastman et al., 2011; Ingram and Mahler, 2013; Khabbazian et al., 2016; O'Meara et al., 2006; Uyeda and Harmon, 2014)

In this work, we propose a method for overcoming the computational complexity of fitting jointly a set of different model types with independent parameter sets to phylogenetically linked comparative data. Our approach relies on a sub-family, hereby denoted $\mathcal{G}_{LInv}$, of the Gaussian phylogenetic models, with the transition density exhibiting the properties that the expectation depends **L**inearly on the ancestral trait value and the variance is **Inv**ariant with respect to the ancestral value. In a related work, we have shown that the likelihood of such models can be calculated in time proportional to the number of nodes in the tree (Mitov et al., 2018). Here, we generalize this fast likelihood calculation algorithm to mixed phylogenetic models over the $\mathcal{G}_{LInv}$-family, which we denote MGPMs. We develop a new algorithm for fast maximum likelihood search of an optimal MGPM fit to a dataset of possibly incomplete measurements from several traits of present day species and/or fossilized specimens, annotating the tips of a time calibrated tree.

A prominent example with a long history in evolutionary biology is the comparative analysis of brain- and body-mass data from mammals (Jerison, 1973; Snell, 1891). In the quest for the origin of intelligence, it has been shown that, in mammals, brain mass has a negative allometric relationship with body-mass, meaning that brain-mass tends to scale at lower proportions with respect to body-mass (Boddy et al., 2012; Jerison, 1973; Montgomery et al., 2010; Snell, 1891). Many studies have compared this allometry between separate mammal clades (see, e.g. (Boddy et al., 2012; Montgomery et al., 2010) and references therein). However, the choice of the groups to be compared in these studies has been driven mainly by the established taxonomic ranking (i.e. order, family, genus) and by the researcher's intuition about which groups "could" be different. Moreover, most of the studies in the past have neglected the phylogenetic relationship between the species within a group, which is a known source of bias in comparative regression analysis (Boddy et al., 2012; Felsenstein, 1985). Here, we show that the MGPM enables a data driven identification of such distinct groups. We have performed an MGPM ML fit to body- and brain-mass data from 629 extant mammal species representative of 21 orders, extracted from the previous works of (Bininda-Emonds et al., 2007) and (Boddy et al., 2012). This revealed a strong statistical support for an MGPM with 12 distinct regimes (11 shifts) comprising both, BM as well as several parameterizations of the OU model. Conditioned on the inferred model parameters, we have reconstructed the ancestral history of the brain-body mass allometry for the past 160 Ma.

This article is organized as follows. In New approaches, we formulate the so-called inter-model shift problem, i.e. the optimization problem aiming at finding the optimal model shifts in a phylogenetic tree with multivariate trait measurements associated with its terminal nodes (tips). Then, we briefly describe our proposed solution based on the MGPM. In Results, we report the analysis of the Mammal data. In Discussion, we provide an interpretation of the results and discuss potential issues of the method and challenges for future work. A detailed description of the methods is provided in Materials and Methods and in Appendix. In Appendix 7.H, we report additional results from a validation test based on simulated data.

7.2 NEW APPROACHES

7.3 THE INTER-MODEL SHIFT PROBLEM

Given number of traits k, a tree $\mathcal{T}$ representing the evolutionary relationship of N species (tips), and a family of k-variate phylogenetic models $\mathcal{M}$, a mixed phylogenetic model on $\mathcal{T}$

is defined as a configuration of shift points and mapped models, $S = \{< 0, m_0 >, < s_1, m_1 > , ..., < s_R, m_R >\}$, where $< 0, m_0 >$ denotes the initial model $m_0 \in \mathcal{M}$ starting from the root (0) and modeling the trait evolution on the descending lineages until reaching a tip from $\mathcal{T}$ or another shift from S; each other shift $< s_i, m_i >$ denotes a point s_i on a branch of $\mathcal{T}$ and a model $m_i \in \mathcal{M}$, assuming the trait values at the point s_i as initial state, and again modeling the evolution on the subtree with root s_i, $\mathcal{T}_{s_i}$, until reaching a tip or a shift (fig. 7.1). We call "shift-point configuration" the set of points where the shifts occur, i.e. $\{0, s_1, ..., s_R\}$. We denote by $\mathcal{S}(\mathcal{T}, \mathcal{M})$ the family of all mixed phylogenetic models over $\mathcal{T}$ and $\mathcal{M}$, with mixed referring to several models on a single tree. The "inter-model shift problem" is the problem of finding the mixed phylogenetic model $S^* \in \mathcal{S}(\mathcal{T}, \mathcal{M})$ that fits "best" to data $\mathcal{X}$ consisting of trait values at the tips of $\mathcal{T}$. We call S^* the best inter-model shift configuration.

Defining "best fit" in the statistical sense is not straightforward, due to the notorious problem of "overfitting" coming along with complex parametric models. In this work, we use the Akaike information criterion (AIC) as a score function penalizing the maximum likelihood (ML) fit of a model, based on the number of free parameters. We note, however, that there is no general agreement on a best scoring function, in particular, for small datasets, where the commonly used AIC and AICc have been shown to be biased towards more complex models (Ho and Ané, 2014b).

7.4 DEALING WITH THE COMPUTATIONAL COMPLEXITY

With a few exceptions (Zwiernik, Uhler, and Richards, 2014), maximizing the likelihood of a mixed phylogenetic model is a multivariate non-convex optimization task involving numerous calculations of the model likelihood for the given tree and data. Furthermore, searching for the best inter-model shift configuration is hard, because the number of possible configurations grows exponential with respect to the number of tips in the tree. Our approach to this complexity is two-fold:

1. THE $\mathcal{G}_{LInv}$ FAMILY OF MODELS. In particular, we restrict $\mathcal{M}$ to a sub-family of the Gaussian phylogenetic models, denoted $\mathcal{G}_{LInv}$. Gaussian phylogenetic models are popular in comparative analysis, because there is a theoretical mapping between microevolutionary forces, such as neutral drift and stabilizing selection, and some Gaussian models, such as BM and OU (Hansen and Martins, 1996). These two models belong to a sub-family of the Gaussian phylogenetic models, $\mathcal{G}_{LInv}$, for which it is possible to calculate the likelihood in time proportional to the size of the tree (Mitov et al., 2018):

Definition 3. *We say that a phylogenetic trait model belongs to the $\mathcal{G}_{LInv}$ family if it satisfies the following*

1. *after branching the traits evolve independently in the two descending lineages,*
2. *the distribution of the trait $\vec{X}$, at time t conditional on the value at time $s < t$ is Gaussian with the mean and variance satisfying*
   a) $\mathrm{E}\left[\vec{X}(t)|\vec{X}(s)\right] = \vec{\omega}_{s,t} + \mathbf{\Phi}_{s,t}\vec{X}(s)$
   (expectation depends linearly on the ancestral trait),
   b) $\mathrm{Var}\left[\vec{X}(t)|\vec{X}(s)\right] = \mathbf{V}_{s,t}$
   (variance is invariant with respect to the ancestral trait, see also fig. 7.2 for an illustration),

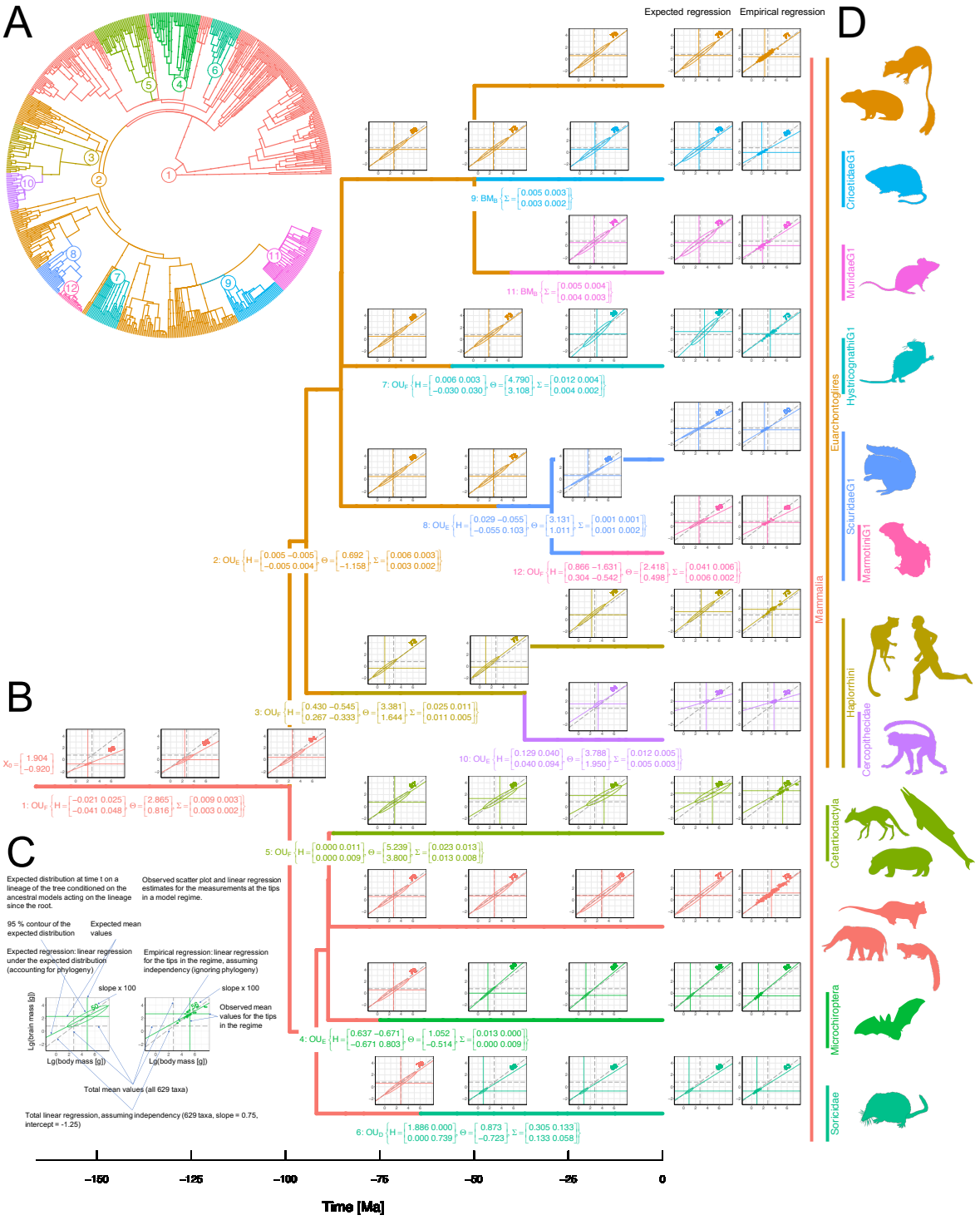


Figure 7.1: **An MGPM reconstruction of the body- and brain-mass evolution in mammals** A: A phylogenetic tree of 629 extant species representative of 21 mammal orders (subsampled from (Bininda-Emonds et al., 2007)). The colours with numbers from 1 to 12 denote the model regimes. B: A pruned (back-bone) variant of the tree in A showing the selected model type and the inferred ML parameters below the shift point for each regime. The plots above the lineages depict the evolution of the trait distribution on each backbone lineage, conditioned on the inferred ML parameters in the MGPM fit. These inferred distributions should be interpreted as the expectation for the corresponding ancestral species under the hypothesis that the inferred MGPM model was the true one. At the tips, this inferred distributions are compared to the empirical trait distributions (scatter plots) localised over the tips belonging to each regime. The ancestral node labels annotating the regimes were provided by Prof. Dr. Jörg Stelling (personal communication). C: Description for the distribution plots in B. D: A visual hint showing some of the mammal species under each regime; artistic images for the species in D (sources and licenses listed in Appendix 7.G).

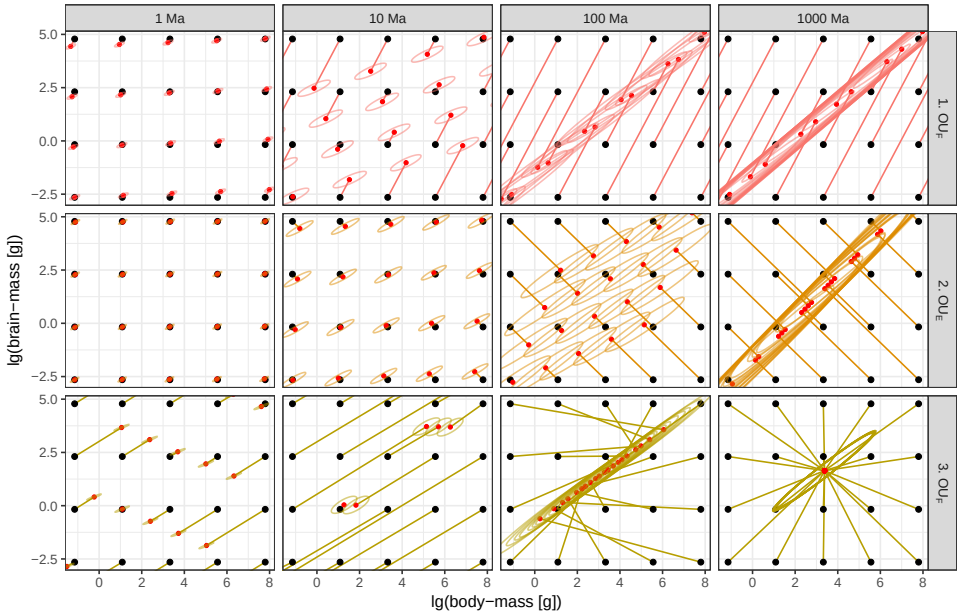


Figure 7.2: **Evolution of brain- and body-mass under selected regimes of the MGPM fit to the mammal data.** Each black point positioned on a regular grid represents the mean value for a given species at some time 0. The corresponding red point (connected by a colored segment) shows how the mean is expected to evolve under the model regime, within a specified time period, Δt. An ellipse around the red-point denotes the 95% contour of the expected normal distribution. If all organisms of the species are divided into isolated sub-populations evolving independently under the corresponding model for a period Δt, at the end of the period, the mean of the sub-populations's mean-values would be approximated by the red point, and approximately 95% of the sub-populations would have their mean values within the ellipse. Rows of panels correspond to different model regimes. Columns from left to right correspond to different time periods. Note that within each panel, the ellipses have the same shape, size and orientation, in accordance with property 2(b), Defn 3. Figures S3-S4) show the corresponding plots for all 12 regimes in the MGPM fit to mammal data.

for some vector $\vec{\omega}_{s,t}$ *and matrices* $\mathbf{\Phi}_{s,t}$, $\mathbf{V}_{s,t}$ *which can depend on* s *and* t *but don't depend on* $\vec{X}(s)$.

In (Mitov et al., 2018), we have proven that for any tree and any phylogenetic model satisfying Defn. 3, it is possible to calculate the likelihood of the model, given multi-trait data for the tips with some tips possibly missing some trait values, through a pruning algorithm, based on analytical integration over the unobserved trait values at the internal nodes of the tree. Here, we have extended this algorithm to support mixed phylogenetic models over the $\mathcal{G}_{LInv}$-family, meaning the type of model may change at inter-model shift points.

2. FAST MODEL SELECTION. As a next step, we developed an algorithm searching for an optimal mixed Gaussian phylogenetic model (MGPM) over a finite subset $\mathcal{M} \subset \mathcal{G}_{LInv}$. To reduce the search space of mixed phylogenetic models, $\mathcal{S}(\mathcal{T}, \mathcal{M})$, we use several "heuristics":

(A) Reducing the number of candidate shift-point configurations:

(A.1) Motivated by the usual lack of statistical power for inferring the precise location or the presence of multiple shifts within a branch (Bastide et al., 2018; Khabbazian et al., 2016), we assume that a shift-point can only occur at the beginning of a branch.

We call the end-node of such a branch a "shift-node". Following this assumption, the number of shift-point configurations is reduced to 2^{M-1}, with M denoting the number of nodes in the tree.

(A.2) We introduce a threshold, q, on the minimal number of tips "visible" from an ancestor shift-node, where visibility means that there is no other shift occurring on a path from the shift-node to any of these tips. Indirectly, this limits the maximum number of shifts to no more than N/q. However, specifying q instead of a maximum number of shifts has a performance benefit, because even a small value of q (e.g. less than 5% of the tree size) effectively reduces the space of possible shift configurations. As a downside, unlike a limit on the maximum number of shifts, specifying q hinders the detection of shifts visible from less than q tips. This is acceptable, given our goal is to detect patterns in big groups of tips, rather than outliers.

(A.3) The best configuration of a given size P (number of shift-nodes) can be obtained from the best configuration of size $P-1$ by adding one of the other possible shifts, nameyly, the one resulting in the best AIC score. This greedy assumption provides a stop criterion for the search procedure, namely, when a configuration has been reached, which's score cannot be improved by inserting a new shift. While not valid in general, this heuristic has proven useful in numerous previous implementations of stepwize AIC optimization on tree models, e.g. (Alfaro et al., 2009; Ingram and Mahler, 2013).

(B) Reducing the number of possible model type mappings for a given shift-point configuration. For each candidate shift-point configuration comprising P shifts, i.e. $P+1$ regimes in total, there are $|\mathcal{M}|^{P+1}$ possible MGPMs. To reduce this number, we implement a heuristic to be applied when P exceeds some user defined number:

(B.1) In the optimal MGPM, the best model from $\mathcal{M}$ associated with a given shift-node i is likely to be the best model fitting to the clade descending from i. Thus, we reduce the set of candidate model types to be mapped to a candidate shift-node, i, to the set $\{M_{current-best}, M_{clade-best}\}$, $M_{current-best}$ denoting the model mapped to i in the currently found best MGPM, $M_{clade-best}$ denoting the best model mapped to i in a single (non-mixed) model fit to that clade.

Using these heuristics, we implemented a parallel recursive clade partition search algorithm solving the inter-model shift problem by returning an (approximate) optimal inter-model shift configuration for a given tree and multivariate trait data at the tips (Appendix 7.A, algorithm 7.1, figs. S1-S2).

7.5 RESULTS

7.5.1 *An MGPM analysis of the brain-body allometry in mammals*

We performed an MGPM fit to the biggest publicly available phylogenetic tree of mammal species with available body- and brain-mass measurements (fig. 7.1A). This is a subtree of 629 extant species with ancestral nodes spanning 166 Ma, which were extracted from the time-calibrated mammal tree published in (Bininda-Emonds et al., 2007). Curated body- and brain-mass data for the species have been provided by previous works (see (Boddy et al., 2012) and references therein). As a preprocessing step aiming to improve the time-resolution

of the identified model-shifts we repeatedly halved, through insertion of singleton nodes, all branches in the tree longer than 16 Ma, until all branches were shorter than 16 Ma. In this way, we obtained an ultrametric tree of 629 tips, and 1063 internal nodes, of which 494 were annotated ancestral bifurcating or multi-furcating nodes (Bininda-Emonds et al., 2007) and 569 were artificially inserted singleton nodes.

The MGPM fit was done over six candidate model types ranging from a model of neutrally and independently evolving traits to a complex model of evolution under selection and causal relationship between the traits. All of these model types were defined as specifications of the BM and the OU models (see Materials and Methods). Further in the text, we denote these model types by BM_A, BM_B, OU_C, OU_D, OU_E, OU_F or by the capital letters $A - F$. The best model fit found by the recursive clade partition algorithm had $AIC^* = -240.57$, log-likelihood $\ell\ell^* = 235.29$ and a total of $p = 115$ parameters, specifying 11 shift-points and a total 12 regimes. As an additional test for possible overfitting, we conducted a MGPM fit over the models $\{A, ..., E\}$ and a fit of the scalar OU model with shifts in the long-term optimum described in (Bastide et al., 2018). The MGPM over the models $\{A, ..., E\}$ produced a fit with 11 regimes (10 shifts) and AIC=-193.43. The scalar OU model produced a fit with 3 regimes (2 shifts) and AIC=+42.36, which was sub-optimal compared to the best single regime OU_F-model (compare AIC value in iteration 1, fig. S1). Based on the significant AIC difference in favor of the MGPM over the models $\{A, ..., F\}$, we retained this model for the interpretation of results.

Figure 7.1B shows a summary of the identified model regimes in the best MGPM fit. Based on the ML estimates for the root node (X_0) and for the model parameters in each regime, we have reconstructed the expected ancestral distributions at fixed time-points along each lineage (fig. 7.1B, see also Appendix 7.D). A visual interpretation of the inferred model parameters for each regime is shown in figs. 7.2, S3-S4.

Under the hypothesis that the inferred MGPM fit is the true model for the data, these distributions represent the expectation for a sample of species evolving independently since the root of the tree. Thus, the least squares regression between the two traits in these distributions is correct with respect to possible correlations caused by shared ancestry. We notice that, with few exceptions (HystricognathiG1, MuridaeG1, CricetidaeG1), these expected regressions agree closely with the empirical regression lines calculated over the species within each regime (see distribution plots at the tips in fig. 7.1B), with empirical meaning that we simply calculated the regression line for the trait values across tips ignoring any phylogenetic relatedness. Conversely, there is a well pronounced difference between the regression lines in different regimes (fig. 7.1B).

Looking back in time, the inferred model suggests the hypothesis that, with slope=0.4, the brain-body-mass allometry has been far more pronounced in the mammal ancestors 160 Ma ago (fig. 7.1B). This slope has increased gradually through time until reaching nowadays levels of ≈ 0.75 for all species in regime 1 (fig. 7.1B). The model shifts are associated with significant changes in the direction and the magnitude of selection forces (figs. 7.1B, 7.2, S3-S4). Regimes 1, 2, 5 and 8 were characterized by the lack of a single long term focal point (figs. S3-S3). In contrast, for regimes 4, 6, 10 and 12, the model suggests convergence to a global mean point within less than 100 Ma (figs. S3-S4). Considering figs. 7.2, S3-S4, it is possible to hypothesize about the direction and strength of selection acting at different points in the phenotype plane. For example, in regime 1, a species weighing 10 kg and with brain-mass 0.3 kg tends to evolve towards smaller masses for both, body and brain (fig. 7.2). Conversely, in regime 2, the same species tends to evolve towards smaller brain but bigger body mass (fig. 7.2).

7.5.2 *Tests on simulated data*

Using the models $A - F$, we conducted a simulation study on random ultrametric and non-ultrametric trees of up to 638 tips. The simulations confirmed that our MGPM fitting procedure correctly identifies clusters in the tree associated with different evolutionary regimes, and accurately discriminates between OU and BM regimes of evolution. We observed a drop in the predictive power with respect to the exact type of model. For example there was a tendency towards favouring simpler versions of the OU model with symmetric selection strength matrix, **H**, to OU models with asymmetric **H** (see Materials and Methods). A detailed report of these simulations is provided in Appendix 7.H, figs. S5-S16, table S1.

7.6 DISCUSSION

The idea of jointly fitting different types of Gaussian models dates back at least since the work of Slater 2013 Slater, 2013, where he measured the statistical support for a shift from an OU to a BM process in the evolution of mammal body size occurring at the end of the Mesozoic (but see Slater, 2014). Later, Clavel et al. implemented a non-pruning algorithm for multivariate likelihood calculation for shifts between BM, OU and the early burst (EB) model of adaptive radiation (Clavel, Escarguel, and Merceron, 2015). These works assume a known point in time where a "global" shift occurs on all lineages of the tree. The more ambitious task of finding "local" inter-model shifts occurring on individual branches has, to our knowledge, not been addressed, although many authors have proposed methods for finding local intra-model shifts in some of the parameters of the OU-model, and under various simplifying assumptions including tree ultrametricity, single trait or independently evolving multiple traits, shared or fixed parameter values between model regimes (e.g. a scalar OU model with a global (scalar diagonal) selection strength matrix and drift matrix for all regimes) (Bastide, Mariadassou, and Robin, 2017; Bastide et al., 2018; Beaulieu et al., 2012; Butler and King, 2004; Eastman et al., 2011; Ingram and Mahler, 2013; Khabbazian et al., 2016; O'Meara et al., 2006; Uyeda and Harmon, 2014).

With respect to the above works, solving the inter-model shift problem over $\mathcal{S}(\mathcal{T}, \mathcal{G}_{LInv})$ should provide a wide modeling possibility for evolutionary biologists. Apart from BM and OU, the $\mathcal{G}_{LInv}$ family includes many popular models of continuous trait evolution, such as BM and OU models with a linear trend in the (long term) mean, jump-enabled BM or OU models of punctuated equlibrium, OU models with separate selection strength and decorrelation rate and EB models (Mitov et al., 2018; Pennell and Harmon, 2013). However, the inclusion of these model types in the MGPM fit should always be subjected to a consideration of the model identifiability in the context of the observed data.

Understanding the identifiability of phylogenetic models in the multiple regime setting is an open problem. This problem consists in the possibility for different shift configurations, model mappings or parameter values to fit equally well to a given tree and data. Previous works have made a progress in understanding the identifiability of single regime or scalar OU models (Bastide et al., 2018; Ho and Ané, 2014b; Khabbazian et al., 2016). For example, (Ho and Ané, 2014b) showed analytically that in a single regime OU model on an ultrametric tree, it is not possible to infer both, the root value $\vec{X}_0$ and the long-term optimum $\vec{\theta}$. This statement may not hold any more in a multiple regime model setting allowing for different evolutionary rates among the regimes. Such a model can be interpreted as a scaling by different factors of the branch lengths in the different regimes, resulting in a non-ultrametric tree, with the

scaling factors being informed by the evolutionary rates. Thus, the joint inference of $\vec{X}_0$ and $\vec{\theta}$ in the MGPM fit on an ultrametric tree could be justifiable in the case of several regimes.

Entangled with the identifiability issue is the problem of quantifying the uncertainty in an MGPM fit. Theoretically, this could be approached in a Bayesian way using reversible-jump Metropolis sampling (Uyeda and Harmon, 2014). However, the high dimensionality of the MGPM, combined with the poor potential for parallelizing the sampling pose a serious computational challenge that goes beyond the scope of this work. Therefore, we caution the reader that our present solution is limited to providing a single (possibly local) optimal point in the vast space of MGPM models. The safest way to interpret this point estimate, as well as any derived quantities such as the ancestral trait distributions (fig. 7.1), is to consider these as evolutionary hypotheses that have to be tested in the light of competing methods or novel data.

7.7 MATERIALS AND METHODS

7.7.1 *Candidate model types for the mammal data*

We defined six candidate models based on the k-variate Ornstein-Uhlenbeck (OU) process, defined by the following stochastic differential equation:

$$d\vec{X}(t) = \mathbf{H}\big(\vec{\theta} - \vec{X}(t)\big)dt + \mathbf{\Sigma}_C dW(t). \tag{7.1}$$

In the above equation, $\vec{X}(t)$ is a k-dimensional real vector, $\mathbf{H}$ is a $k \times k$-dimensional eigen-decomposable real matrix, $\vec{\theta}$ is a k-dimensional real vector, $\mathbf{\Sigma}_C$ is a $k \times k$-dimensional real positive definite matrix and $W(t)$ denotes the k-dimensional standard Wiener process. Seen as a branching stochastic process, where each branching event gives rise to two independent instances of the process starting from the value of $\vec{X}$ at the branching point, eq. 7.1 satisfies Defn. 3 (Mitov et al., 2018). Specifically, the elements $\vec{\omega}_{s,t}$, $\mathbf{\Phi}_{s,t}$ and $\mathbf{V}_{s,t}$ from property 2 in Defn. 3 are given by (Mitov et al., 2018):

$$\begin{aligned}
\vec{\omega}_{s,t} &= \left(\mathbf{I} - \mathrm{Exp}\big(-(t-s)\mathbf{H}\big)\right)\vec{\theta} \\
\mathbf{\Phi}_{s,t} &= \mathrm{Exp}(-(t-s)\mathbf{H}) \\
\mathbf{V}_{s,t} &= \textstyle\int_0^{t-s} \mathrm{Exp}(-v\mathbf{H})(\mathbf{\Sigma}_C\mathbf{\Sigma}_C^T)\mathrm{Exp}(-v\mathbf{H}^T)dv
\end{aligned} \tag{7.2}$$

Biologically, $\vec{X}(t)$ denotes the mean values of k continuous traits in a species at a time t from the root, the parameter $\mathbf{\Sigma} = \mathbf{\Sigma}_C\mathbf{\Sigma}_C^T$ defines the magnitude and shape of the momentary fluctuations in the mean vector due to genetic drift, the matrix $\mathbf{H}$ and the vector $\vec{\theta}$ specify the trajectory of the population mean through time. When $\mathbf{H}$ is the zero matrix, the process is equivalent to Brownian motion and the parameter $\vec{\theta}$ is irrelevant. When $\mathbf{H}$ has strictly positive eigenvalues, the population mean converges in the long term towards $\vec{\theta}$, although the trajectory of this convergence can be complex (see figs. 7.2, S3, S4). In all parametrizations, we restrict $\mathbf{H}$ to have non-negative eigenvalues - a negative eigenvalue of $\mathbf{H}$ transforms the process into repulsion with respect to $\vec{\theta}$, which, while biologically plausible, is not identifiable in a ultrametric tree. The six candidate models are specified below:

- BM_A ($\mathbf{H} = 0$, diagonal $\mathbf{\Sigma}$): BM, uncorrelated traits;
- BM_B ($\mathbf{H} = 0$, symmetric $\mathbf{\Sigma}$): BM, correlated traits;

- OU_C (diagonal $\mathbf{H}$, diagonal $\mathbf{\Sigma}$): OU, uncorrelated traits;
- OU_D (diagonal $\mathbf{H}$, symmetric $\mathbf{\Sigma}$): OU, correlated traits, but simple (diagonal) selection strength matrix;
- OU_E (symmetric $\mathbf{H}$, symmetric $\mathbf{\Sigma}$): An OU with non-diagonal symmetric $\mathbf{H}$ and non-diagonal symmetric $\mathbf{\Sigma}$;
- OU_F (asymmetric $\mathbf{H}$, symmetric $\mathbf{\Sigma}$): An OU with non-diagonal asymmetric $\mathbf{H}$ and non-diagonal symmetric $\mathbf{\Sigma}$;

7.7.2 *Implementation*

The technical details of the recursive clade partition search, the AIC score, the model parametrizations and the calculation of the expected distributions and linear regression coefficients in the mammal data are described in Appendix, sections 7.A-7.E. The pruning algorithm for fast likelihood calculation of MGPM models over the $\mathcal{G}_{LInv}$-family was implemented within the R-package PCMBase (https://github.com/venelin/PCMBase), using internal calls to its companion Rcpp extension PCMBaseCpp (https://github.com/venelin/PCMBaseCpp) Mitov et al., 2018 and the SPLITT library for tree traversal (https://github.com/venelin/SPLITT) (Mitov and Stadler, 2017b). Gradient descent likelihood optimization with multiple calls to optim from random starting parameters was implemented within the R-package OptimMCMC (https://github.com/venelin/OptimMCMC). The recursive clade partition algorithm was implemented in the R-package PCMFit (https://github.com/venelin/PCMFit). The scripts and the data for the mammal analysis and the simulations were implemented in the R-package TestPCMFit (https://github.com/venelin/TestPCMFit). These packages rely on numerous third party libraries listed in Appendix 7.F.

7.8 ACKNOWLEDGEMENTS

V.M. and T.S. thank ETH Zürich for funding. KB's research is supported by Vetenskapsrådets grant no. 2017–04951. The authors wish to thank Prof. Dr. Jörg Stelling for providing the analyzed mammal data including the taxonomic labels for the internal nodes of the tree and Dr. Joelle Barido-Sottani for useful discussions.

APPENDIX

APPENDIX

7.A RECURSIVE CLADE PARTITION SEARCH FOR AN OPTIMAL MGPM

In algorithm 7.1, we provide a pseudo-code description of the recursive clade partition search. To understand how the algorithm works, it may be useful to follow the search path for the optimal MGPM fit to the mammal tree shown on figs. S1-S2. We note that, a shift occurs at the beginning of a branch leading to a so called "shift-node". We call "selected nodes" the nodes that have been selected as shift-nodes in the current best MGPM.

The algorithm starts with a ML fit of each model type to each clade of at least $q = 20$ tips, including the entire tree. The results of these model fits are stored in a data-table, which is used as a source for proposals of initial parameters in subsequent MGPM ML fits. Then, the algorithm initiates a queue of "partition root", starting with the root of the tree. Partition root are equivalent to already selected shift-nodes. In each iteration of the main loop (line 12, algorithm 7.1), the partition root at the head of the queue is taken, and an attempt is made to improve the current best MGPM model by inserting a shift at one of its descendants. We call "candidates" the descendant nodes of a partition root, which have not been cut-out by (i.e. do not descend from) a previously selected shift-node and which satisfy the requirement that, after placing a shift on their corresponding branch, no regime (color) in the resulting tree would have less than q tips.

Each panel on figs. S1-S2 shows the state at the beginning of a main-loop iteration. This state comprises the iteration number (number in parentheses), the AIC, log-likelihood and total number of parameters for the current best MGPM, the currently selected shift-nodes (colored points), the partition root (a colored point with a number equal to the iteration number), the candidate nodes (grey points) and the candidate model types for both, the selected and the candidate nodes (sets of capital letters in braces above the corresponding nodes). During the iteration, a maximum likelihood fit is performed for all MGPMs formed by adding one candidate (grey) node to the set of selected (colored) nodes and for all possible model mappings on this node and the currently selected shift-nodes. Note that this is a greedy step following heuristic A.2 in the main text. As an option it is possible to relax this heuristic by considering combinations of up to a user-specified number of candidate nodes. However, this would considerably slow down the search. The set of possible model mappings for a configuration of shift-nodes can be formed as the Cartesian product of all candidate model types taken for each node. This, however, would result in an exponentially growing number of possible mappings. Thus, a reduction is made by using the heuristic B.1 in the main text. For the partition root, the heuristic B.1 is neglected and all possible model types are considered. For the other nodes, up to 2 model types are considered (the best model fit to the clade starting at the node vs the model assigned to the node in the current best-fit). This effectively reduces the number of possible model mappings, although, in the worst case this number would still be exponential, i.e. in the order of 2^s, where s denotes the number of shift-nodes in the shift configuration. If during the main-loop iteration the AIC has been improved by inserting a new shift-node, this shift-node, together with the partition root are inserted at the end of the queue, so that a further partition from these nodes can be explored in a next iteration. The algorithm ends when the partition queue gets empty.

Algorithm 7.1 : Recursive clade partition search for an optimal MGPM

Input :
$\mathcal{T}$: a timed tree with M nodes of which N are tips;
$\mathcal{X} \in (\mathbb{R} \cup \{NA, NaN\})_{k \times N}$: data for k traits associated with the tips, missing values or non-existing traits allowed;
$\mathcal{M} \subset \mathcal{G}_k$, $|\mathcal{M}| < \infty$: a finite set of k-variate Gaussian phylogenetic models;
MLE: $\bigcup_{i \in \{0,N+1,\ldots,M-1\}} \mathcal{S}_i(\mathcal{T}_i, \mathcal{M}) \longrightarrow \{< \ell\ell^*, \Theta^* >\}$: a maximum likelihood estimator getting as input a mixed Gaussian phylogenetic model (a shift configuration and model mapping) on (a subtree of) $\mathcal{T}$ and returning the corresponding maximum likelihood, $\ell\ell^*$, and point estimate, Θ^*, of model parameters contained in S;
SCORE: $\{< S, \ell\ell >\} \longrightarrow \mathbb{R}$: a scoring function, penalizing a maximum likelihood value $\ell\ell$ based on the complexity, e.g. the number of free parameters of the model S;
Output : A quasi-optimal MGPM, $S^* \in \{\mathcal{S}(\mathcal{T}, \mathcal{M})\}$, with respect to *SCORE*.

Data :
TableFits: a table with columns *tree*, *model*, Θ and q, containing (an encoded/compressed version of) the tree, the MGPM the parameter-values and the penalized score for all MLEs produced during the search;
QueuePartitionRoots: a first-in-first-served list (queue) of the nodes used as clade-partition roots during the search;
S^*: the current MGPM on $\mathcal{T}$ with best score;

Step 1. Initialization. *Fit each individual model to each clade in* $\mathcal{T}$.
foreach $i \in \{0, N+1, ..., M-1\}$ **do**
  foreach $m \in \mathcal{M}$ **do**
    $S_{i,m} \longleftarrow \{< i, m >\}$;
    $< \ell\ell^*_{i,m}, \Theta^*_{i,m} > \longleftarrow MLE(S_{i,m}; \mathcal{T}_i, \mathcal{X}_i, \mathcal{M})$;
    $q^*_{i,m} \longleftarrow SCORE(S_{i,m}, \ell\ell^*_{i,m})$;
    Add to *TableFits* $\langle tree = \mathcal{T}_i, model = S_{i,m}, \Theta = \Theta^*_{i,m}, q = q^*_{i,m} \rangle$;

Step 2. Recursive clade-partition search for the optimal MGPM on $\mathcal{T}$.
Step 2.1. *Initialize QueuePartitionRoots with root-node and the best individual model fit to* $\mathcal{T}$ *found in TableFits.*
Add to QueuePartitionRoots $< 0 >$;
$S^* \longleftarrow$ {model in *TableFits* with the best score on the whole tree};
`// Main loop`
while *QueuePartitionRoots is not empty* **do**
  Step 2.2. *Get the node at the head of the queue: this node is the partition root for the iteration.*
  $j \longleftarrow PopFrontElement(QueuePartitionRoots)$;
  Step 2.3. *Extract the subtree of* $\mathcal{T}$ *containing all tips descending from j without an intermediate node from* S^* *on their path to j.*
  $\mathcal{T}'_j \longleftarrow ExtractClade(\mathcal{T}, j)$;
  foreach $l \in Nodes(S^*) \backslash \{j\}$ **do**
    if $l \in Nodes(\mathcal{T}'_j)$ **then**
      $\mathcal{T}'_j \longleftarrow RemoveClade(\mathcal{T}'_j, l)$;
  $PartitionNodes \longleftarrow Nodes(\mathcal{T}_j)'$;
  Step 2.4. *Make a list of all shift configurations including* $Nodes(S^*)$ *and a node from PartitionNodes.*
  $P \longleftarrow \phi$;
  foreach $p \in PartitionNodes$ **do**
    $P \longleftarrow P \bigcup \{Nodes(S^*) \cup \{p\}\}$;
  Step 2.5. *Restrict the set of candidate models for each node in* $S^* \cup PartitionNodes \backslash \{j\}$ *to at most two models; try all models for the node j.*
  foreach $l \in S^* \cup PartitionNodes \backslash \{j\}$ **do**
    $\mathcal{M}_l \longleftarrow$ {model assigned to l in S^*} $\cup$ {best model fit to clade l};
  $\mathcal{M}_j \longleftarrow \mathcal{M}$;
  Step 2.6. *MLE fits to all shift configurations in P and possible model mappings using* $\mathcal{M}_l, l \in P$
  foreach $Sp \in P$ **do**
    foreach $Sm \in \prod_{l \in Sp} \mathcal{M}_l$ **do**
      $S \longleftarrow \{< Sp, Sm >\}$;
      $< \ell\ell^*_S, \Theta^* > \longleftarrow MLE(S; \mathcal{T}, \mathcal{X}, \mathcal{M})$;
      $q^* \longleftarrow SCORE(S, \ell\ell^*_S)$;
      Add to *TableFits* $\langle tree = \mathcal{T}, model = S, \Theta = \Theta^*, q = q^* \rangle$;
  Step 2.7. *If step 2.6 has found a fit with a better score than the score of* S^*, *update* S^* *and add its nodes to the queue.*
  if $TableFits[tree == \mathcal{T}, Min(q)] < SCORE(S^*, \ell\ell_{S^*})$ **then**
    $S^* \longleftarrow BestModel(TableFits[tree == \mathcal{T}])$;
    Add to QueuePartitionRoots $Nodes(S^*)$;
return S^*;

7.A.1 *Likelihood optimization*

We used calls of the R-function optim specifying the L-BFGS-B algorithm for optimizing the likelihood of each candidate MGPM. To reduce the risk of getting stuck in local optima, multiple runs have been performed starting from different locations in the parameter space. These starting locations were specified as follows:

- Clade fits: for the initial step of the algorithm, in which each model type is fit to each clade of not less than q tips, the likelihood of the MGPM was evaluated at a large number (in this case 200'000) of parameter points drawn at random from a uniform distribution defined by user-specified limits (see Parameter limits in the Model parametrizations section 7.C). Then the points were sorted in order of decreasing likelihood and optim was run for the top 400 points.
- Main loop fits: for the main loop MGPM fits, we implement a similar procedure as for the clade fits, but with reduced number of likelihood evaluations (4000) and 10 optim calls. The starting locations have been chosen from a mixture of randomly drawn parameters and slightly modified (jittered) optimum points from the clade fits for each shift-node and mapped model type.

7.A.2 *Parallel execution*

We implemented parallel execution of the nested foreach loops in step 1 (line 2) and step 2.6 (line 28) in algorithm 7.1. Parallelization was implemented within the PCMFit R-package via calls to the R-packages foreach ("foreach: Foreach Looping Construct for R"), iterators ("iterators: Iterator Construct for R") and doMPI ("doMPI: Foreach Parallel Adaptor for the Rmpi Package"). The MGPM fit for both the mammal data and the simulated data (Appendix section 7.H) was performed on the Euler cluster managed by the HPC team at ETH Zurich. For the analysis of the mammal dataset the search algorithm finished within 24 hours, running on 300 cores (299 MPI worker nodes). For the analysis of the simulated data, the search algorithm finished within 24 hours for 159 out of 192 datasets, running on 100 cores (99 MPI worker nodes). The remaining 33 datasets were only for big trees ($N = 638$ tips, see section 7.H) and took between two and four days.

7.B CALCULATING THE AIC OF A MGPM ML FIT

For a ML fit of the MGPM model, the Akaike information criterion is given by the formula

$$AIC = -2\ell\ell^* + 2p \tag{S1}$$

where $\ell\ell^*$ denotes the maximum log-likelihood and p denotes the number of the parameters. For the MGPM on a fixed tree and data, we define p as the total number of numerical model parameters, that is, the initial trait vector, $\vec{X}_0$ together with the parameters for each model regime, plus $[2 * (R - 1) + 1]$, where R denotes the number of regimes. In this way, every shift counts as 2 added parameters (shift location and mapped model), and a single parameter is counted for the model-type in the root-regime.

7.C MODEL PARAMETRIZATIONS

7.C.1 *Transformations for the matrix parameters* $\mathbf{\Sigma}$ *and* $\mathbf{H}$

We used transformations on the matrix parameters $\mathbf{\Sigma}$ and $\mathbf{H}$ to prevent the likelihood optimization from hitting on invalid parameter values (e.g. non-symmetric or non-positive-definite matrix $\mathbf{\Sigma}$, or negative-definite matrix $\mathbf{H}$). We note that the same techniques described briefly below have been used in other OU implementations, e.g. (Bartoszek et al., 2012; Clavel, Escarguel, and Merceron, 2015).

For the unit-time variance-covariance matrices, $\mathbf{\Sigma}$, which are symmetric positive definite by definition, we used the parametrization:

$$\mathbf{\Sigma} = \mathbf{\Sigma}_C \mathbf{\Sigma}_C^T, \tag{S2}$$

where $\mathbf{\Sigma}_C$ denotes the upper-triangular Choleski factor of $\mathbf{\Sigma}$ (Bartoszek et al., 2012; Clavel, Escarguel, and Merceron, 2015). Note that by specifying positive values for the diagonal elements of $\mathbf{\Sigma}_C$, we guarantee that $\mathbf{\Sigma}$ is positive definite. Further, the fact that $\mathbf{\Sigma}_C$ is triangular guarantees the symmetry of $\mathbf{\Sigma}$.

For the OU selection strength matrices $\mathbf{H}$, which we require to have non-negative eigenvalues without necessarily being symmetric (note that negative eigenvalues result into repulsion from the $\vec{\theta}$), we used the Schur parametrization following (Clavel, Escarguel, and Merceron, 2015). Specifically, we define a $k \times k$-dimensional matrix $\mathbf{H}_S$ as follows:

- the upper triangle of $\mathbf{H}_S$, excluding the diagonal, specifies $k(k-1)/2$ rotation angles for Givens rotations (Golub and Van Loan, 2012) to obtain a $k \times k$-dimensional orthoganal matrix $\mathbf{Q}$;
- the lower triangle of $\mathbf{H}_S$ including the diagonal defines a $k \times k$ triangular matrix T.

Then, $\mathbf{H}$ is obtained from $\mathbf{Q}$ and $\mathbf{T}$ as follows (Bartoszek et al., 2012; Clavel, Escarguel, and Merceron, 2015):

$$\mathbf{H} = \mathbf{Q}\mathbf{T}^T\mathbf{Q}^T \tag{S3}$$

The matrix $\mathbf{H}$ calculated in this way has all of its eigenvalues equal to the elements on the diagonal of $\mathbf{H}_S$ (Bartoszek et al., 2012; Clavel, Escarguel, and Merceron, 2015). Thus, by restricting the diagonal of $\mathbf{H}_S$ to non-negative values, we guarantee that $\mathbf{H}$ will have all of its eigenvalues non-negative. Further, if $\mathbf{H}_S$ is diagonal, then so is be the matrix $\mathbf{H}$; if $\mathbf{H}_S$ is upper triangular, then $\mathbf{T}$ is diagonal and the resulting matrix $\mathbf{H}$ is symmetric. Finally, if $\mathbf{H}_S$ is a full matrix, i.e. neither diagonal nor triangular, then the resulting matrix $\mathbf{H}$ is asymmetric.

7.C.2 *Parameter limits*

Since we used a the L-BFGS-B algorithm for gradient-descent optimization (Byrd et al., 1995), we need to specify limits for the model parameters. We did this as follows:

- $0.0 \leqslant \mathbf{\Sigma}_{C,ii} \leqslant 1$, $i \in \{1,2\}$ for all model types;
- $0.0 \leqslant \mathbf{\Sigma}_{C,12} \leqslant 1$ for all model types;
- $0.0 \leqslant \mathbf{H}_{S,ii} \leqslant 10$, $i \in \{1,2\}$ for all OU model types;

- $-10.0 \leqslant \mathbf{H}_{S,12} \leqslant 10.0$ for the OU_E model type (keeping $\mathbf{H}_{S,21} = 0$ to ensure symmetry of the transformed matrix $\mathbf{H}$);
- $-10.0 \leqslant \mathbf{H}_{S,ij} \leqslant 10.0$, $i \neq j \in \{1,2\}$ for the OU_F model type;
- $0.0 \leqslant \theta_1 \leqslant 7.8$ according to the range of lg-body-mass in grams in the mammal dataset;
- $-1.2 \leqslant \theta_2 \leqslant 3.8$ according to the range of lg-brain-mass in grams in the mammal dataset.

7.D CALCULATING EXPECTED TRAIT DISTRIBUTIONS UNDER THE MGPM

We use the fact that, under an MGPM model of the evolution of k traits along a tree $\mathcal{T}$, the expected distribution of the trait values at any time point, i, on any branch of $\mathcal{T}$ is a k-variate Gaussian distribution. The mean k-vector and the $k \times k$ variance-covariance matrix of this distribution are functions of the initial (root) trait vector, $\vec{X}_0$ and the model parameters and branch lengths for the sequence of regimes on the path from the root to i. These functions are calculated by applying Defn. 3 in the following recursive fashion:

1. Node 0 (the root of $\mathcal{T}$) is associated with a k-variate Dirac's δ with infinite density over the root-value and 0 density elsewhere:

$$\begin{aligned} \mathrm{E}\left[\vec{X}_0\right] &= \vec{X}_0, \\ \mathrm{Var}\left[\vec{X}_0\right] &= [0]_{k\times k}. \end{aligned} \tag{S4}$$

2. For any other point i, let j be the closest ancestor of i and t and s be their corresponding time distances from the root. Then the expected distribution of the trait vector at i, $\vec{X}_i$ is a k-variate Gaussian with mean and variance given by:

$$\begin{aligned} \mathrm{E}\left[\vec{X}_i\right] &= \vec{\omega}_{s,t} + \mathbf{\Phi}_{s,t}\,\mathrm{E}\left[\vec{X}_j\right], \\ \mathrm{Var}\left[\vec{X}_i\right] &= \mathbf{V}_{s,t} + \mathbf{\Phi}_{s,t}\,\mathrm{Var}\left[\vec{X}_j\right]\mathbf{\Phi}_{s,t}^T, \end{aligned} \tag{S5}$$

   where $\vec{\omega}_{s,t}$, $\mathbf{\Phi}_{s,t}$ and $\mathbf{V}_{s,t}$ are defined as in Defn. 3.

7.E ORDINARY LEAST SQUARES REGRESSIONS

For calculating the linear regression lines of lg-brain-mass on lg-body-mass (fig.7.1), we use the fact that for a bivariate normal distribution of two variables x and y with mean vector $\vec{\mu} = [E(x), E(y)]^T$ and variance covariance matrix $\mathbf{V} = \begin{bmatrix} \sigma^2(x) & \sigma(x,y) \\ \sigma(x,y) & \sigma^2(y) \end{bmatrix}$, the linear regression of y on x, i.e. the linear model $y = a + bx + \epsilon$, has ordinary least squares (OLS) estimates for the slope (b) and intercept (a) given by the equations:

$$\begin{aligned} b &= \sigma(x,y)/\sigma^2(x) \\ a &= E(y) - bE(x). \end{aligned} \tag{S6}$$

Using eq. S6, we calculated the intercept and the slope of the OLS regressions of lg-brain-mass on lg-body-mass in the mammal data as follows (see also fig. 7.1):

(a) First, we calculated the expected distributions of the two traits under the inferred MPGM fit in the different backbone lineages (fig. 7.1) at seven past points in time located at regular intervals of 27 Ma, starting from -162 Ma and ending at the present time. For the calculation, we used the expected mean vector, $\vec{\mu} = [E(lg-body-mass), E(lg-brain-mass)]^T$ and variance-covariance matrix $\mathbf{V}$ under the MGPM fit (eqs. S4 and S5, Appendix 7.D).

(b) We additionally calculated an empirical regression based on the tip trait values, ignoring the phylogenetic relationship in the tree. In particular, the empirical measurements of the two traits for the tips in the tree stratified by inferred MGPM regime. We note that these OLS regressions assume independency of the tips, neglecting the correlation due to shared ancestry between the species in each regime. This is a known source of bias (Felsenstein, 1985) motivating the use of PCMs. To do the OLS calculation, we used eq. S6, plugging in the empirical mean and variance covariance matrices. We cross-validated the resulting values for the coefficients with the slope and intercept obtained from calling the R-function lm.

(c) Third, we calculated the empirical measurements of the two traits for all 629 tips in the tree, using the empirical mean and variance-covariance matrices. The resulting OLS estimates matched the values used as a reference for the calculation of encephalization quotient (EQ) in (Boddy et al., 2012). Again, we stress that this regression line (dashed grey regression lines on fig. 7.1) are calculated without accounting for the phylogenetic correlation between the tips.

7.F THIRD PARTY LIBRARIES

The software packages accompanying this article rely on a number of third party libraries: ape (Paradis, Claude, and Strimmer, 2004), Armadillo (Sanderson and Curtin, 2016), expm ("expm: Matrix Exponential, Log"), mvtnorm (Genz and Bretz, 2009), data.table (Dowle et al., 2014), ggplot2 (Wickham, 2009), ggtree (Yu et al., 2017), ggimage ("ggimage: Use Image in 'ggplot2'"), foreach ("foreach: Foreach Looping Construct for R"), doMPI ("doMPI: Foreach Parallel Adaptor for the Rmpi Package"), iterators ("iterators: Iterator Construct for R"), digest ("digest: Create Compact Hash Digests of R Objects"), Rcpp (Eddelbuettel, 2013). Additional tools used to generate the simulation data and to produce the figures and tables include phytools (Revell, 2011), cowplot ("cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2' [R package cowplot version 0.9.3]"), knitr (Xie, 2017), rmarkdown (Allaire et al., 2014) and xtable ("xtable: Export Tables to LaTeX or HTML").

7.G IMAGES USED IN FIG. 7.1

For fig. 7.1D, monochrome images were downloaded from http://phylopic.org. The vectorized or raster images were re-colored using Microsoft Office or Adobe Illustrator. The images are licensed either under the Public Domain Dedication 1.0 license (hereby abbreviated as PDD 1.0) available at `https://creativecommons.org/publicdomain/zero/1.0/`, or under the Creative Commons Attribution-ShareAlike 3.0 Unported license available at `http://creativecommons.org/licenses/by-sa/3.0/` (hereby abbreviated as CCASAU 3.0). Below, we list the images used, their authors, phylopic.org-ids and licenses:

- Cricetidae by Natasha Vitek: 81930c02-5f26-43f7-9c19-e9831e780e53, PDD 1.0;

- Hystricognathi by Zimices: 8d4497e7-a1b4-49e8-9e02-23db5f9aa37c, CCASAU 3.0;
- Muridae by Daniel Jaron: 92989e35-4e68-4a2d-b3a2-191ba9da671a, PDD 1.0;
- Haplorrhini (1) by Gareth Monger: 24230275-1bfa-4ec2-a946-ca1ececdf216, CCASAU 3.0;
- Haplorrhini (2, Homo sapiens sapiens) by T. Michael Keesey 2b4c32f6-99d0-43ba-9180-8013aa5bccd2, PDD 1.0;
- Cercopithecidae (uncredited), eccbb404-c99f-41f9-8785-01a7f57f1269, PDD 1.0;
- Marmotini by T. Michael Keesey 61440e34-7d24-4607-8479-2708ac45663f, PDD 1.0;
- Cetartiodactyla (1, Artiodactyla) by T. Michael Keesey 407f51d5-aa40-4e71-a5a7-7a6d6f328b5d, PDD 1.0;
- Cetartiodactyla (2, Cetacea) by Scott Hartman e68270c1-3091-4aee-92ae-51341a40e94a, PDD 1.0;
- Cetartiodactyla (3, Hippopotamus amphibius) by Jan A. Venter, Herbert H. T. Prins, David A. Balfour & Rob Slotow (vectorized by T. Michael Keesey) , 6336f90c-8f02-48f5-94d1-1d85c0100473, CCASAU 3.0;
- Microchiroptera by Yan Wong, 18bfd2fc-f184-4c3a-b511-796aafcc70f6, PDD 1.0;
- Soricidae by Becky Barnes, 822c549b-b29b-47eb-9fe3-dc5bbb0abccb, PDD 1.0;
- Sciuridae by Catherine Yasuda, 5ebe5f2c-2407-4245-a8fe-397466bb06da, PDD 1.0;
- Feliformia (uncredited), ec56fa32-947b-4f0c-976b-c456132f2d6e, PDD 1.0;
- Diprotodontia by Michael Scroggie, f5592cab-cc61-4aab-b1dd-fba7cd2df7c9, PDD 1.0;
- Euarchonta by T. Michael Keesey (after Joseph Wolf), 88a07585-846a-405d-9195-c15c010e7443, PDD 1.0;
- Elephantidae by T. Michael Keesey, a15244a4-ecaa-4891-b870-31e5c8d9b5b3, PDD 1.0;

7.H SIMULATIONS

We performed a benchmark on two-trait data simulated on ultrametric and non-ultrametric birth-death trees of small (N=318) and big (N=638) sizes (figs. S5-S6).

The trees were generated using calls to the function pbtree from the R-package phytools as follows:

- treeFossilSmall <- pbtree(n=200, scale=1, b = 1, d = 0.4) : generated a non-ultrametric tree of size $N = 318$ (the size depends on the random generator seed).
- treeExtantSmall <- pbtree(n=318, scale=1, b = 1, d = 0.4, extant.only = TRUE) : generated an ultrametric tree of size $N = 318$.
- treeFossilBig <- pbtree(n=374, scale=1, b = 1, d = 0.4) : generated a non-ultrametric tree of size $N = 638$ (the size depends on the random generator seed).

- treeExtantBig <- pbtree(n=638, scale=1, b = 1, d = 0.4, extant.only = TRUE) : generated an ultrametric tree of size $N = 638$.

To match the time-scale of the mammal tree, we rescaled the branch-lengths in the trees so that their total height would be equal to 166.2. This allowed to perform trait simulation and ML-inference on the same scale for the parameters as in the analysis of the mammal data.

For each tree, we assigned two shift-point configurations as follows:

- 1 shift point, i.e. 2 regimes;
- 7 shift points, i.e. in 8 regimes.

For each shift-point configuration, we generated 4 random model type mappings drawing random models from the set $\{BM_A, BM_B, OU_C, OU_D, OU_E, OU_F\}$ as specified in the main text. For each model mapping, we generated three random MGPMs drawing their parameter sets from uniform distributions as follows:

- $0.05 \leqslant \mathbf{\Sigma}_{C,ii} \leqslant 0.5$, $i \in \{1,2\}$ for all model types;
- $0.0 \leqslant \mathbf{\Sigma}_{C,12} \leqslant 0.2$ for all model types;
- $0.1 \leqslant \mathbf{H}_{S,ii} \leqslant 4.0$, $i \in \{1,2\}$ for all OU model types;
- $-4.0 \leqslant \mathbf{H}_{S,12} \leqslant 4.0$ for the OU_E model type (keeping $\mathbf{H}_{S,21} = 0$ to ensure symmetry of the transformed matrix $\mathbf{H}$);
- $-4.0 \leqslant \mathbf{H}_{S,ij} \leqslant 4.0$, $i \neq j \in \{1,2\}$ for the OU_F model type;
- $3.0 \leqslant \theta_1 \leqslant 6.0$ for all OU model types;
- $2.0 \leqslant \theta_2 \leqslant 4.0$ for all OU model types;

Fixing the starting point to $X_0 = (1.0, -1.0)^T$, for each randomly drawn parameter-set, we simulated two random data-sets, using the function PCMSim from the package PCMBase (Mitov et al., 2018). This resulted in a total of $2 \times 2 \times 2 \times 4 \times 3 \times 2 = 192$ simulated data-sets (figs. S7-S14).

For each simulated data-set we ran an MGPM fit over the models $\{BM_A, ..., OU_F\}$ specifying the same boundaries for the parameters as in the mammal data analysis (see Model parametrizations).

The recursive clade partition algorithm was run for each simulation with the same settings as for the mammal data analysis, except for the number of parallel CPU cores which was reduced from 300 to 100. All inferences were run on the Euler HPC.

7.H.1 *Performance assessment*

To evaluate the performance of a MGPM fit to a given simulated tree and data, we define six criteria. The first criterion is how the AIC score of the best fit identified during the search compares against the AIC score calculated for the true model used to generate the data on the tree. If the AIC score of the best fit is bigger (worse) than the AIC of the simulated model, we know for sure that the optimum of the AIC surface could not be found during the search. Very likely the fit has been stuck in a local optimum, away from the true model. Conversely, if the found AIC is smaller (better) than the true model's AIC, there is a chance that the

global optimum has been found. Still, this does not imply that the true model parameters are located in the same valley of the AIC surface.

We defined five additional criteria, each one representing a question with a binary (positive or negative) answer that can be asked either for every pair of nodes in the tree or for every branch in the tree. We compare the answers to these questions given by the best MGPM fit against the known true answers. The true positive rate (tpr, also known as *sensitivity*) is calculated as the proportion of actual positive cases (pairs of nodes or branches, depending on the criterion) that are correctly identified as positive by the fit. The false positive rate (fpr, also known as $1 - specificity$) is calculated as the proportion of negative cases that are wrongly identified as positive by the fit. The perfect fit to a given data and tree has fpr=0 and tpr=1 for each criterion. The worse fit to a given data and tree has fpr=1 and tpr=0 for each criterion. Equality between the tpr and the fpr for some criterion corresponds to a random guess. Hence, we evaluate the performance of a fit for a given criterion as the point (fpr, tpr) located inside the unit square. The five criteria are listed below:

1. Cluster: for each pair of nodes in the tree (internal and tip nodes), we ask if the branches leading to these nodes evolve under the same regime (i.e. have the same color). The test is positive if the two branches do belong to the same regime and negative otherwise.

2. OU process: for each branch in the tree, we ask whether it evolves under an OU model, i.e. one of the models C, D, E, F. Note that this criterion can not be evaluated for model mappings where none of the mapped model types was among C, D, E, F (all negative, so impossible to calculate tpr) or all of the model types were among C, D, E, F (all positive, so impossible to calculate fpr).

3. Correlated traits: for each branch in the tree, we ask whether the regime assigned to that branch supports correlated traits, that is, the model type mapped to that regime is among B, D, E, F. Similar to criterion 2, this criterion could not be evaluated for model mappings where none of the mapped model types was among B, D, E, F or all of the model types were among B, D, E, F.

4. NonDiagonal H: for each branch in the tree, we ask whether its regime has a non-diagonal matrix $\mathbf{H}$, that is, the model mapped to that regime is among the model types E and F. Similar to criteria 2 and 3, this criterion could not be evaluated for model mappings where none of the mapped model types was among E, F, or all of the model types were among E, F.

5. Asymetric H: for each branch in the tree, we ask whether its regime has an asymetric matrix $\mathbf{H}$, that is, the model type mapped to that regime is F. Same considerations as above apply when all or none of the mapped model types are equal to F.

Figs. S15-S16 show the performance for the 192 fits and table S1 summarizes the results over different groupings of simulations. We comment on these results in the following paragraphs.

SMALL ULTRAMETRIC TREES WITH TWO REGIMES. We observed better AIC for all inferred models (nearly all labels white in the first row of fig. S15). There was very good performance with respect to criteria 1, 2 and 3 (excluding two simulations with fpr=1 for criterion 3). For criterion 1 (Cluster), we observe fpr$\approx$0 for nearly all simulations, but there is a tendency towards tpr<1. This indicates a tendency of the best fit to have one or two more regimes than the true number. For criteria 4 and 5 (NonDiagonal H, Asymmetric H)

several of the MGPM fits were located at the diagonal, while all the others were located near the ideal point (fpr=0,tpr=1), suggesting that the signal for accurately detecting these two properties could depend strongly on the actual simulated data, as well as the model fit.

SMALL NON-ULTRAMETRIC TREES WITH TWO REGIMES. The presence of labels in black in the second line of fig. S15 shows some tendency for the best found MGPM to be sub-optimal with respect to the true model. The performance for the different criteria was slightly worse for criterion 2, where several simulations had fpr>0, indicating a bias in favour of the OU process. Due to the random assignment of model types, there were no simulations containing both, regimes for correlated and regimes for non-correlated traits, hence the empty panel for criterion 3 (Correlated traits). Numerous labels away from the ideal point (fpr=0,tpr=1) for criteria 4 and 5 showed that making the tree non-ultrametric did not improve the detectability of OU models with NonDiagonal or Asymmetric H matrices - observing fpr≈0 while tpr<1 shows that, in both cases, the search for an optimal fit has favoured the more parsimonous model, i.e. C or D instead of the true E or F for criterion 4 and C, D or E instead of the true F for criterion 5.

SMALL ULTRAMETRIC TREES WITH EIGHT REGIMES. With all labels in black on the third row in fig. S15, it is clear that the search algorithm has failed to find the valley of the true model in the MGPM space. Despite that, we still observe overall good performance for criteria 1 and 2 and for some of the simulations in criterion 3. For criteria 4 and 5, though, most of the labels were far away from the ideal with a concentration around (fpr=0, tpr=0).

SMALL NON-ULTRAMETRIC TREES WITH EIGHT REGIMES. In this case, we observe a better performance of the search algorithm (predominantly white labels). While the performance is relatively similar fo the ultrametric case, there is a tendency for the white labels to cluster closer to the ideal point, whereas black labels group closer to the red diagonal (criteria 2, 3, 4).

BIG TREES. With big trees (fig. S16), the performance was similar to the performance in small trees, with several exceptions: the segregation between white and black labels was more pronounced, in particular for criterion 1, ultrametric trees with 2 regimes; the AIC for the best fit for ultrametric trees with 8 regimes was overly better than the AIC of the true model; the AIC for the best fit for non-ultrametric trees with 8 regimes was worse than the AIC of the true model.

7.I SUPPLEMENTARY FIGURES

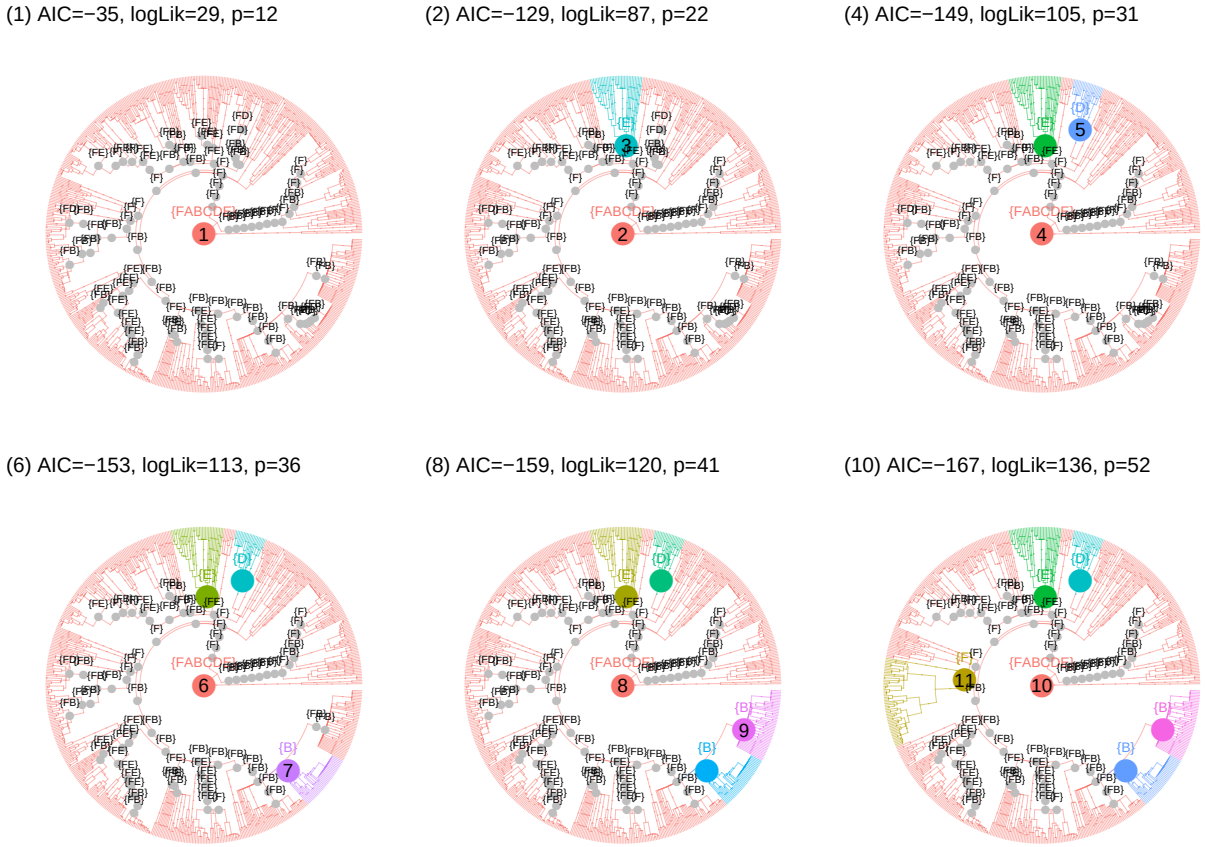
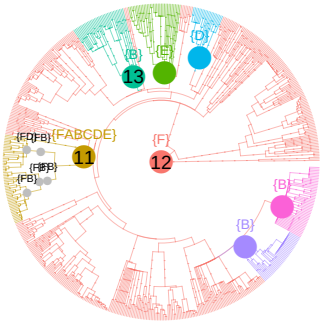
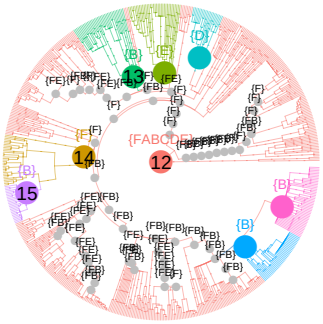


Figure S1: **Search path of the recursive clade-partition algorithm in the mammal tree.** As initialization step, each model-type is fit to each clade not smaller than a user-defined threshold, q (here, $q = 20$). Each panel denoted by a number in parentheses (i) describes iteration i of the main loop (line 11 in algorithm 7.1). The coloured node with a number i is the partition root for the iteration. ColoNodes in grey represent the potential shift points - these are descendants from the partition root, which have not been "cut out" by a shift and have at least q descendants, themselves. Letters in braces denote the candidate model-types for each shift-node. For the partition root (i), these are all model-types; for every other node, this is the set $\{XY\}$, where X is the model-type assigned to the node in the best fit on the entire tree found so far, and Y is the best model-type fit to the node's clade during the initial step.

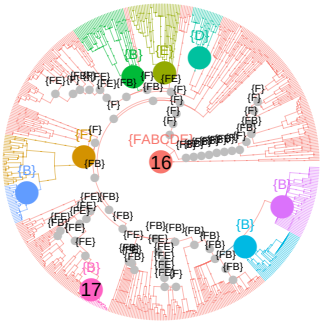
(11) AIC=−188, logLik=151, p=57



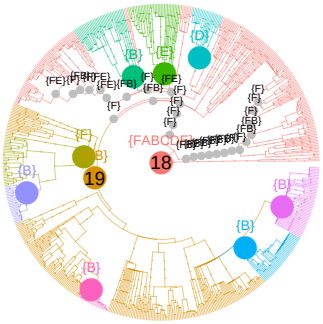
(12) AIC=−189, logLik=156, p=62



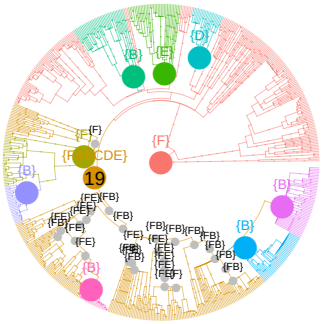
(16) AIC=−202, logLik=168, p=67



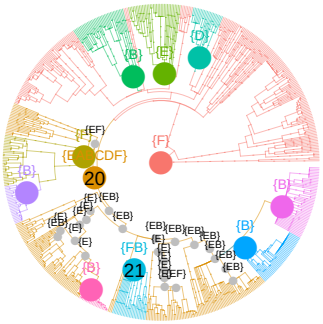
(18) AIC=−222, logLik=189, p=78



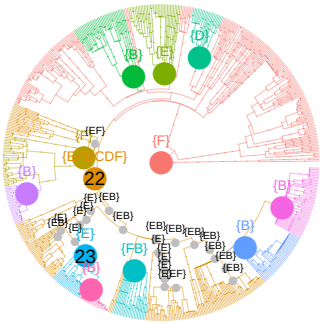
(19) AIC=−222, logLik=189, p=78



(20) AIC=−234, logLik=205, p=88



(22) AIC=−237, logLik=217, p=98



FINAL: AIC=−241, logLik=235, p=115

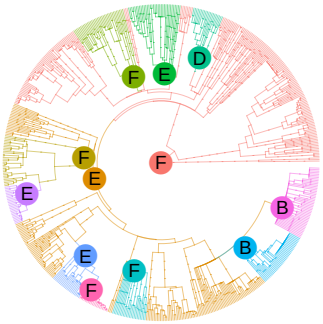


Figure S2: **Search path of the recursive clade-partition algorithm in the mammal tree.**

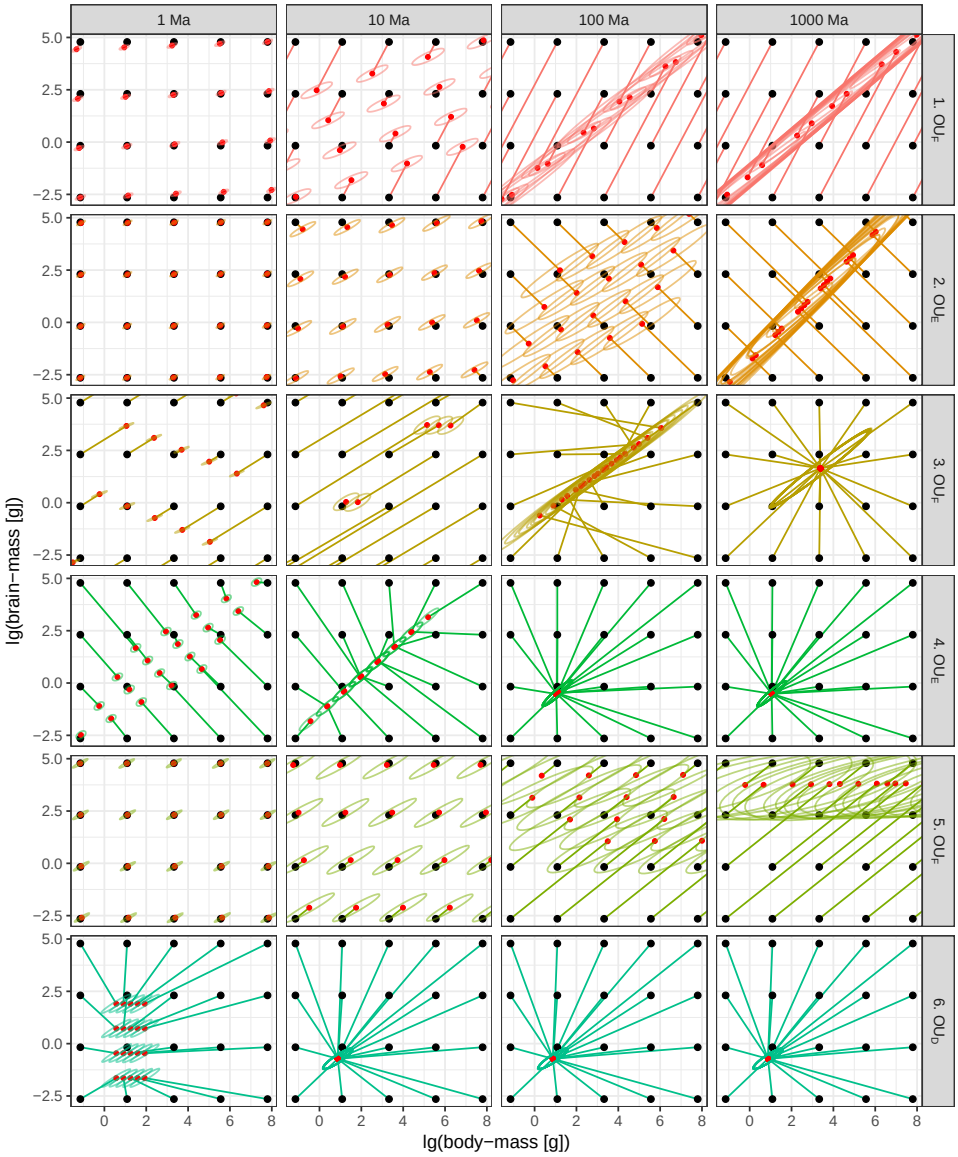


Figure S3: **Evolution of lg-brain and lg-body mass in mammals regimes 1 to 6.** See also legend for fig. 7.2 in the main text.

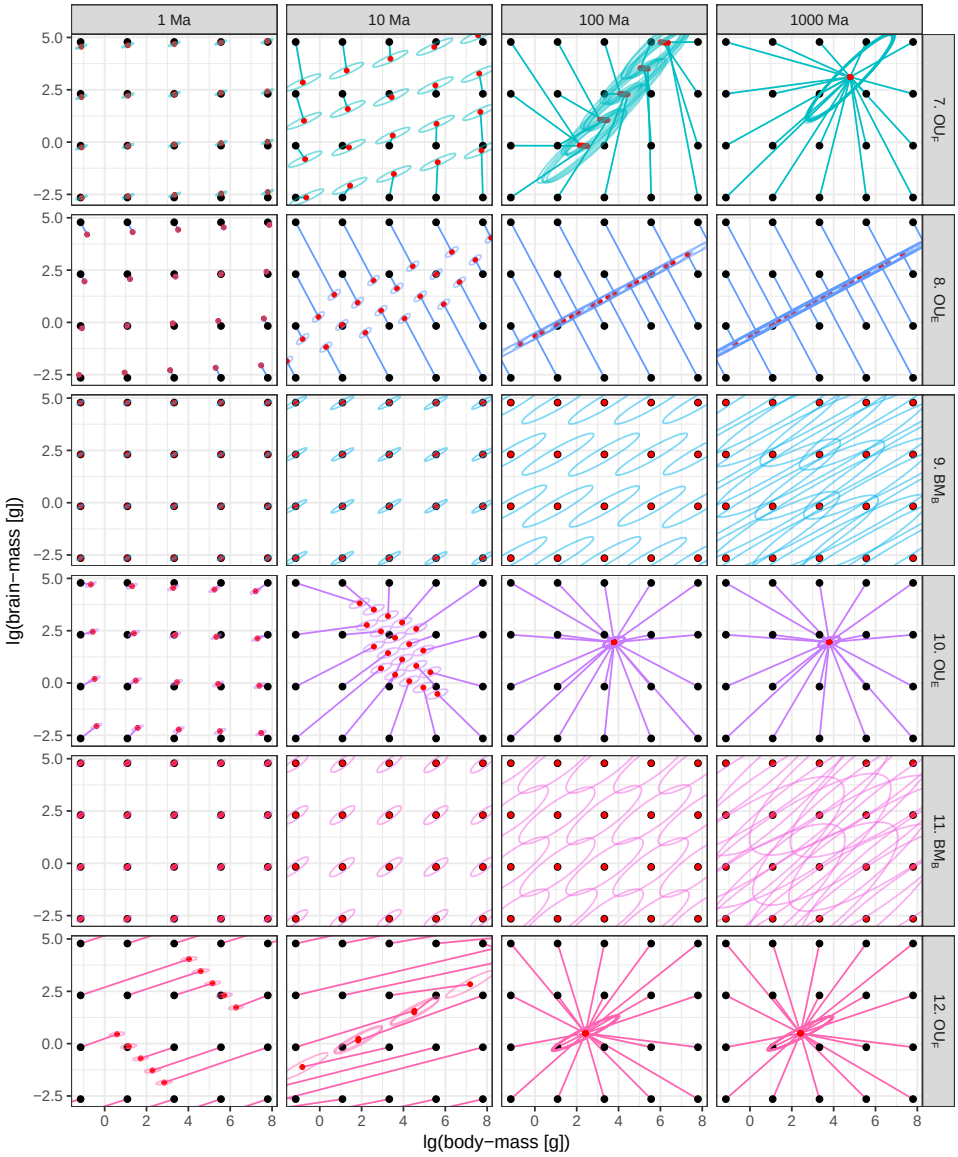
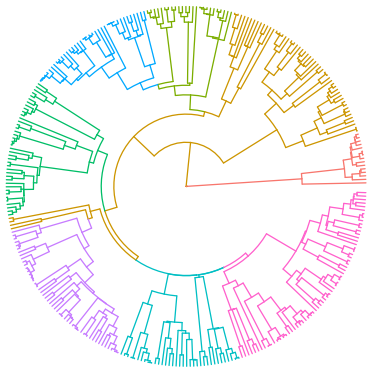


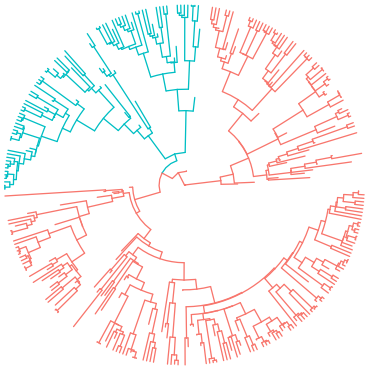
Figure S4: **Evolution of lg-brain and lg-body mass in mammals regimes 7 to 12.** See also legend for fig. 7.2 in the main text.

A. ultrametric / 2 regimes

B. ultrametric / 8 regimes



C. non−ultrametric / 2 regimes



D. non−ultrametric / 8 regimes

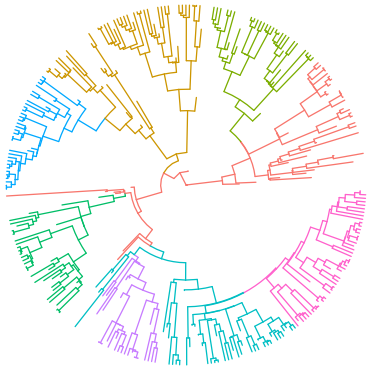


Figure S5: **Simulated small birth-death trees.**

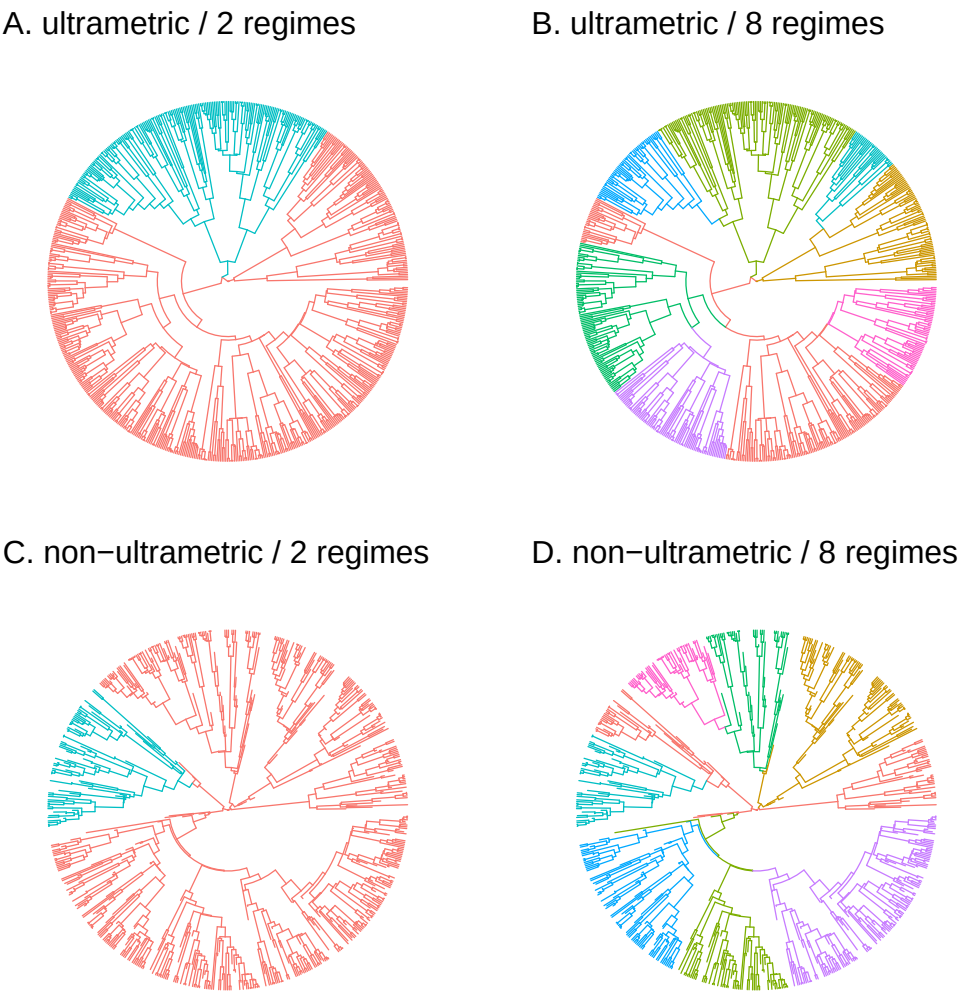


Figure S6: **Simulated big birth-death trees.**

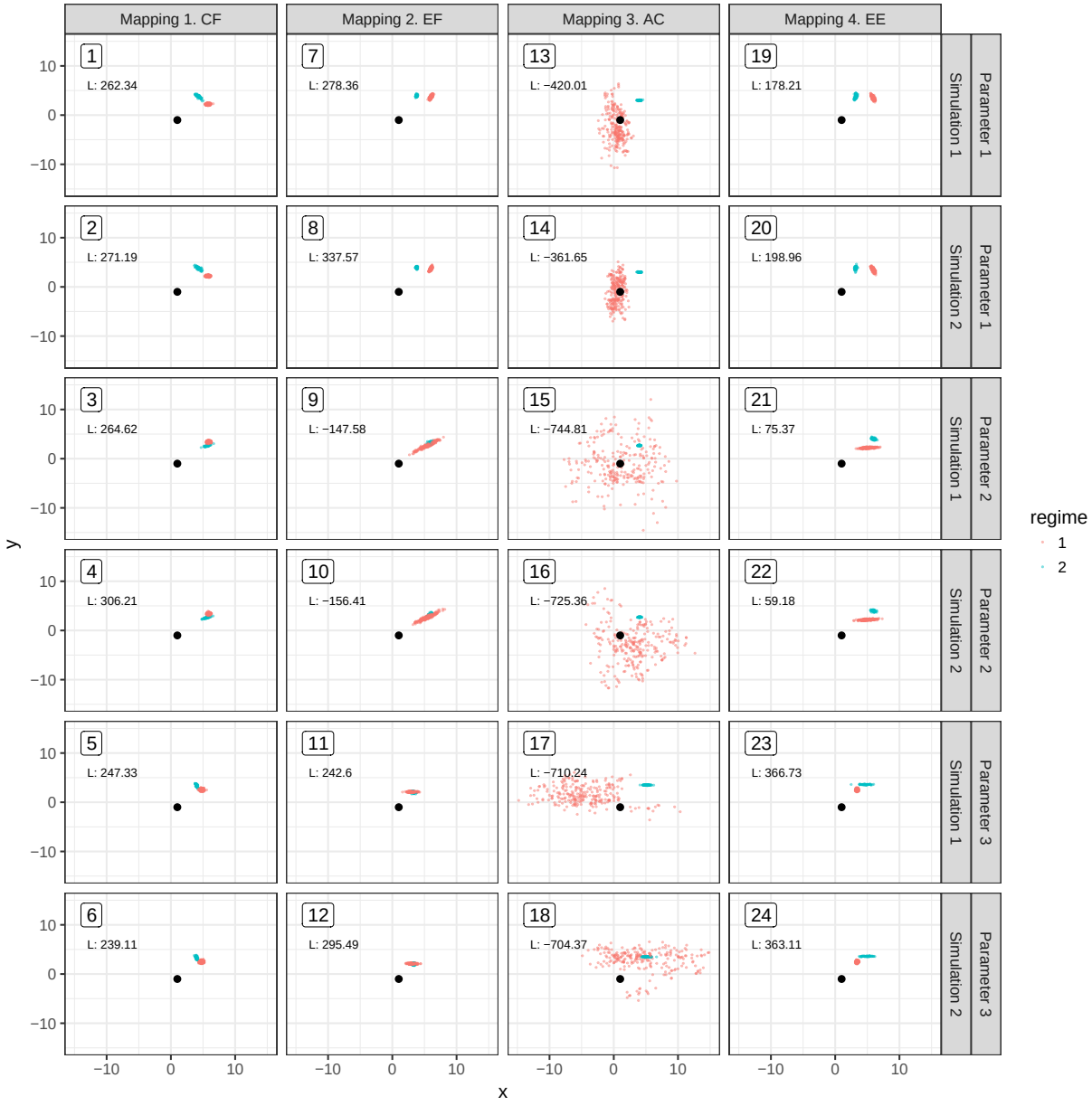


Figure S7: **Scatter plots of the simulated data-sets for small ultrametric trees with 2 regimes.** In each panel, each colored point represents the values of the two simulated traits (x and y) for one tip in the tree. The number label in the top-left corner of each panel denotes the identifier of the simulation, which can be used to look-up the simulated data in the testData_t2 data.table of the accompanying package TestPCMFit. The label denoted by capital letter L denotes the log-likelihood of the data evaluated under the true model. A black point denotes the starting trait value for each simulation. For non-ultrametric trees (figs. S11-S14), the distance from the root is denoted by the transparency of the colored points, paler points denoting closer distance to the root.

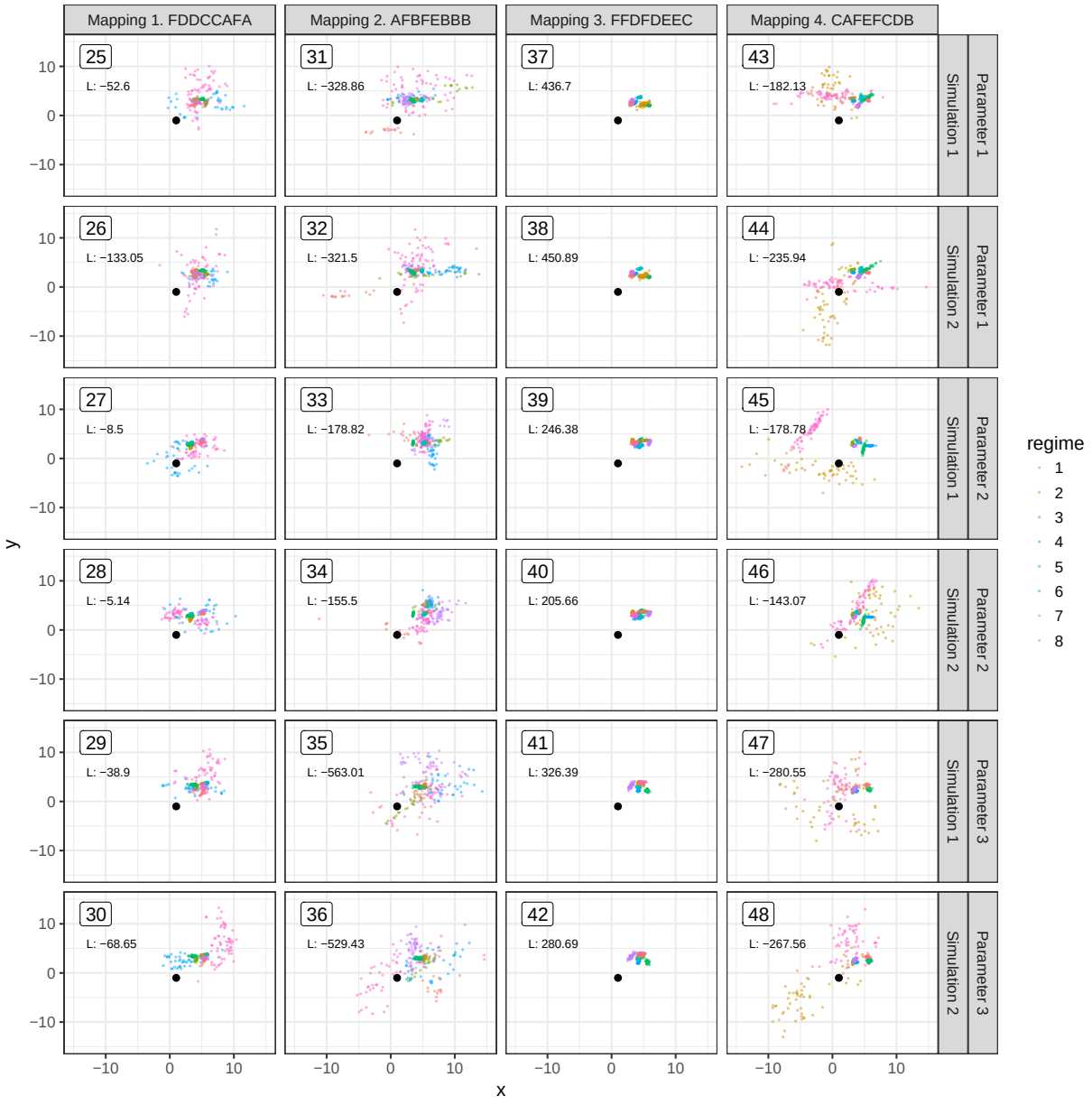


Figure S8: **Scatter plots of the simulated data-sets for small ultrametric trees with 8 regimes.** See legend for fig. S7.

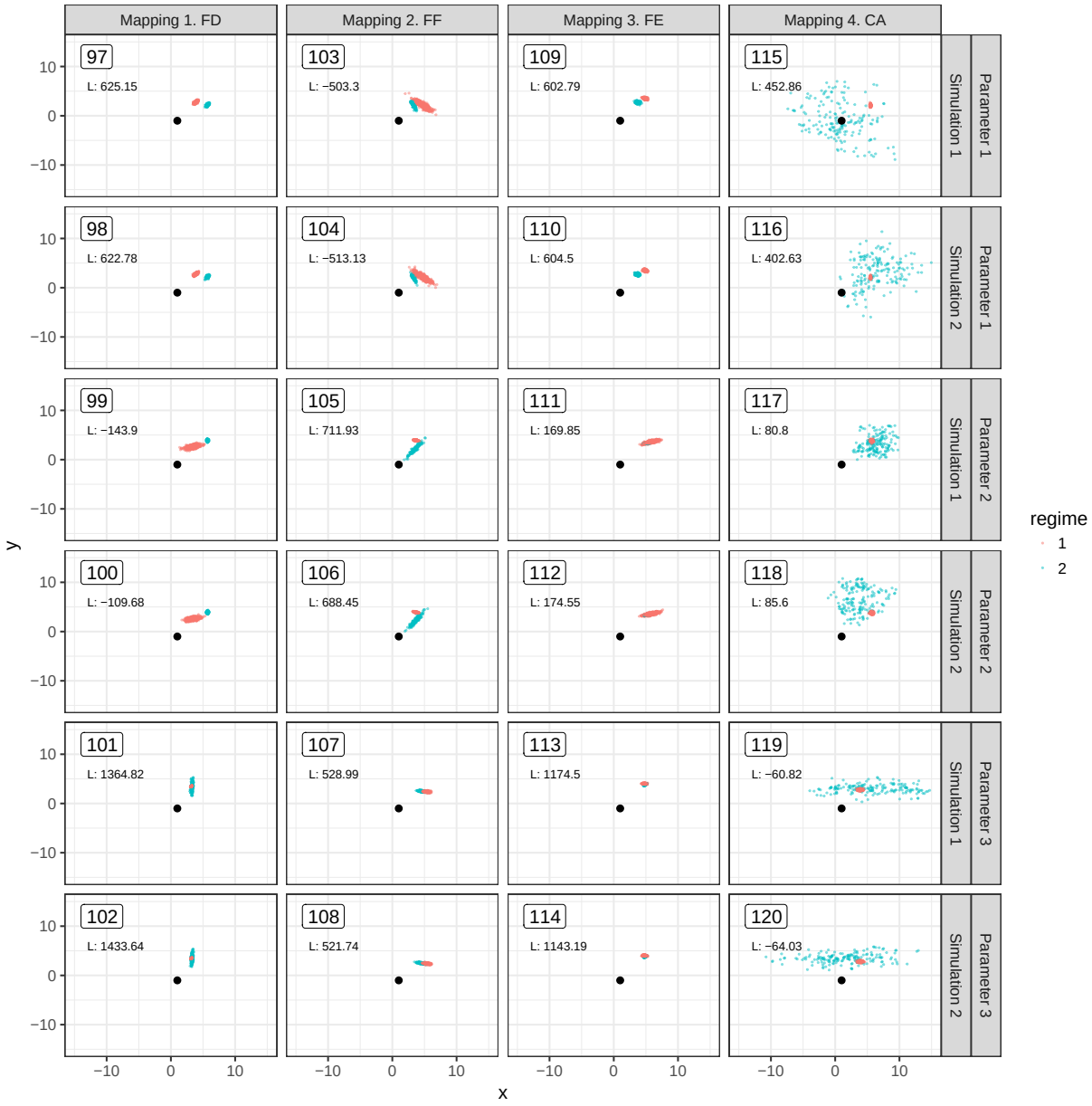


Figure S9: **Scatter plots of the simulated data-sets for big ultrametric trees with 2 regimes.** See legend for fig. S7.

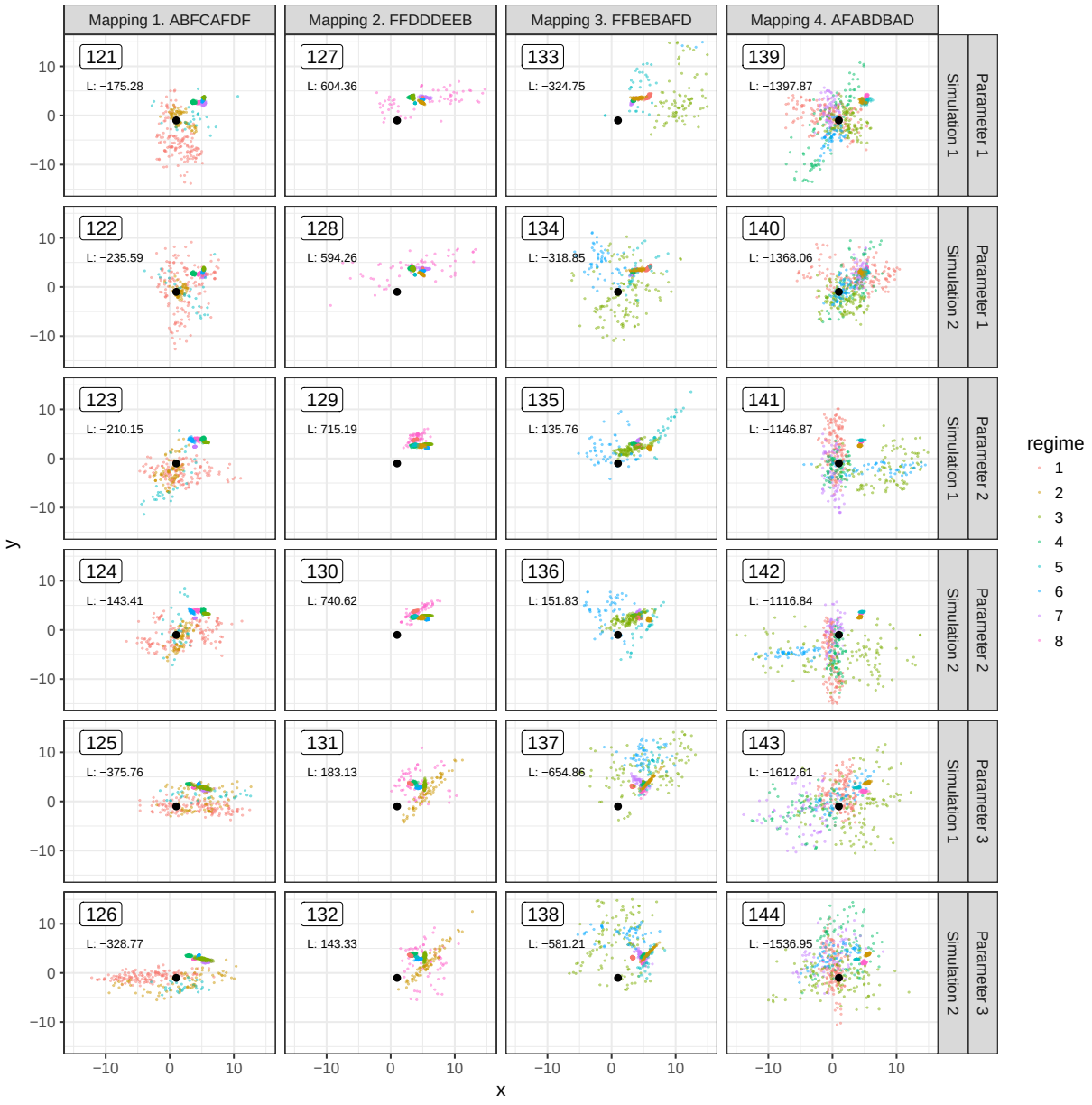


Figure S10: **Scatter plots of the simulated data-sets for big ultrametric trees with 8 regimes.** See legend for fig. S7.

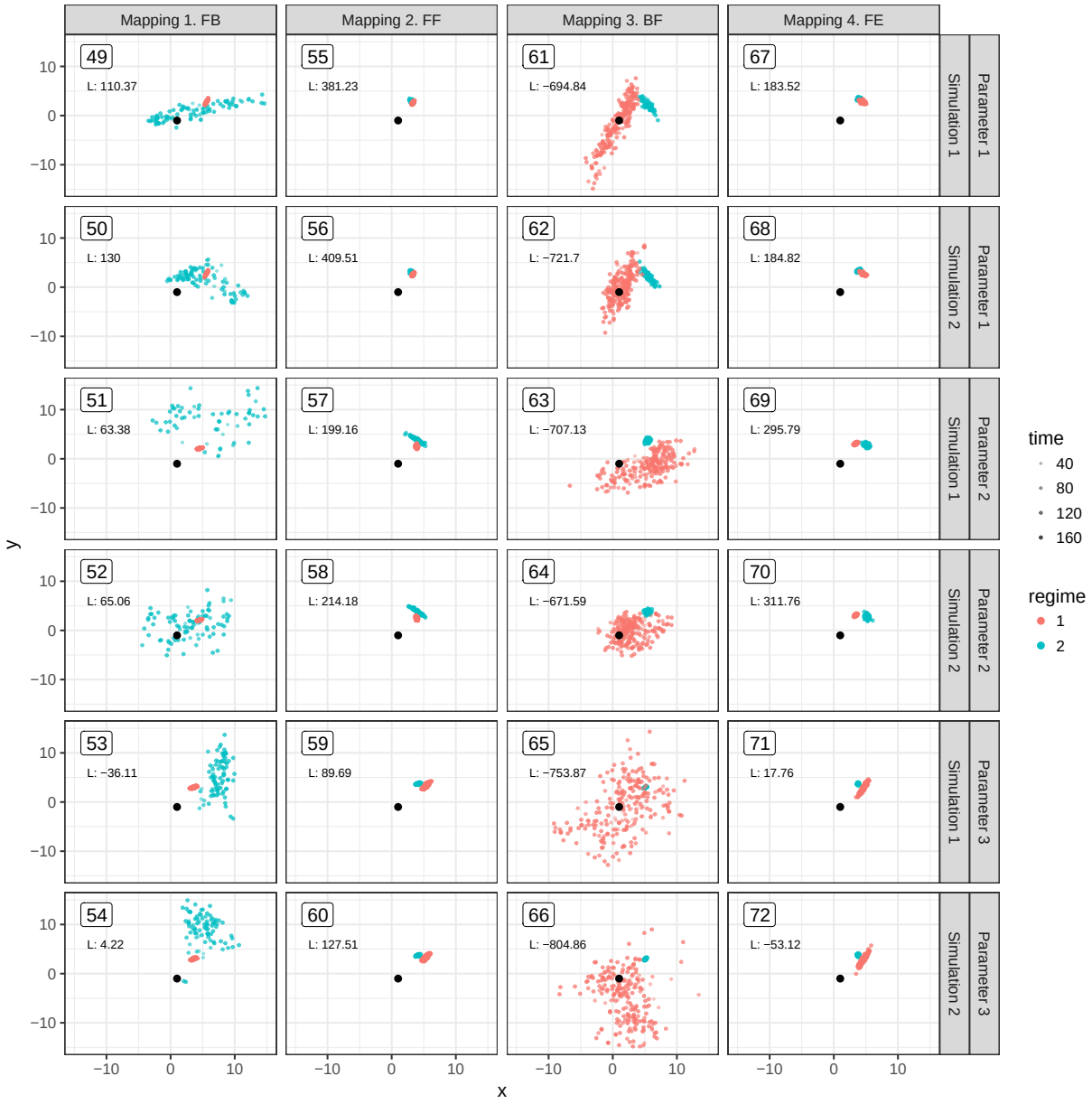


Figure S11: **Scatter plots of the simulated data-sets for small non-ultrametric trees with 2 regimes.** See legend for fig. S7.

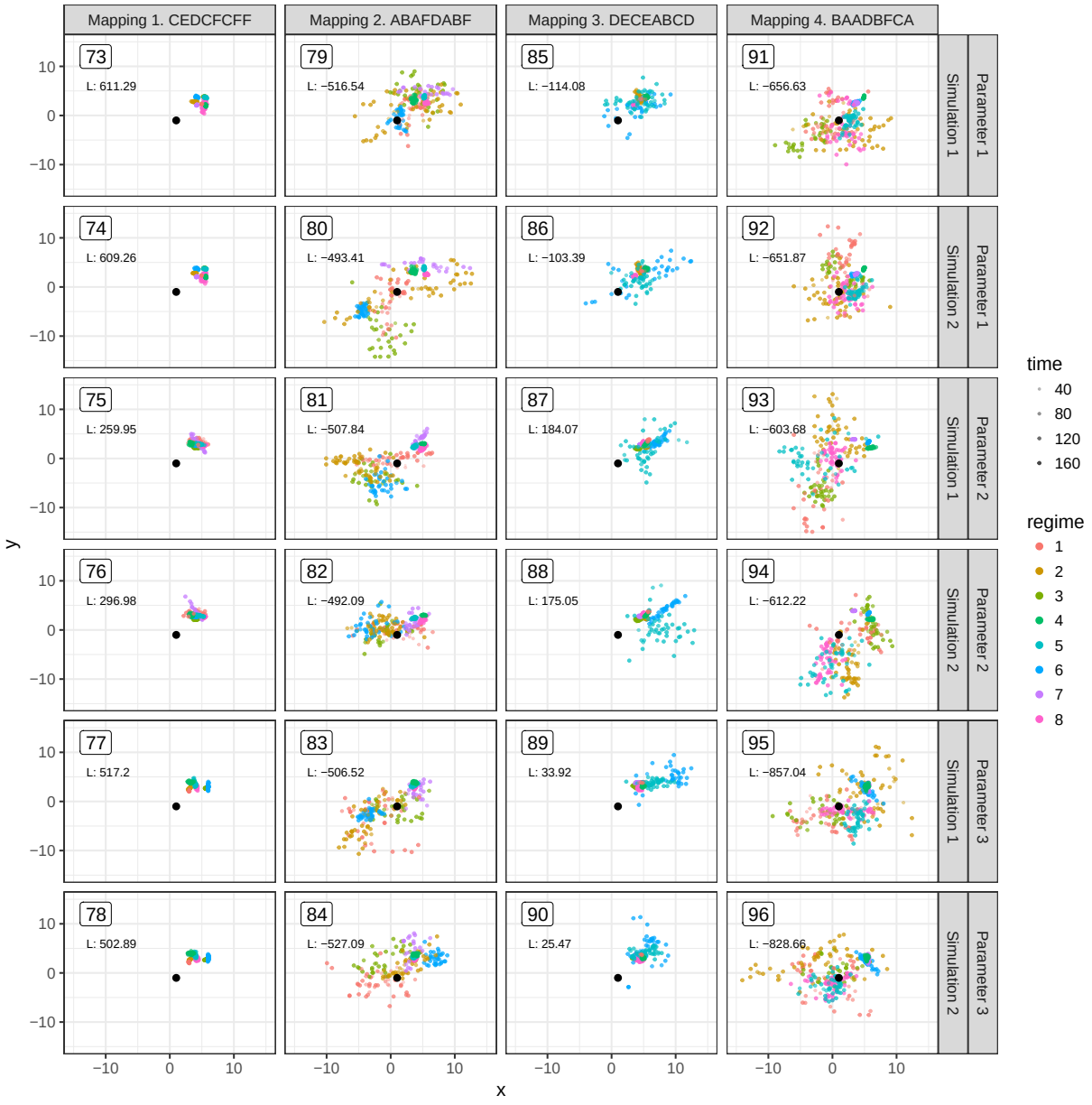


Figure S12: **Scatter plots of the simulated data-sets for small non-ultrametric trees with 8 regimes.** See legend for fig. S7.

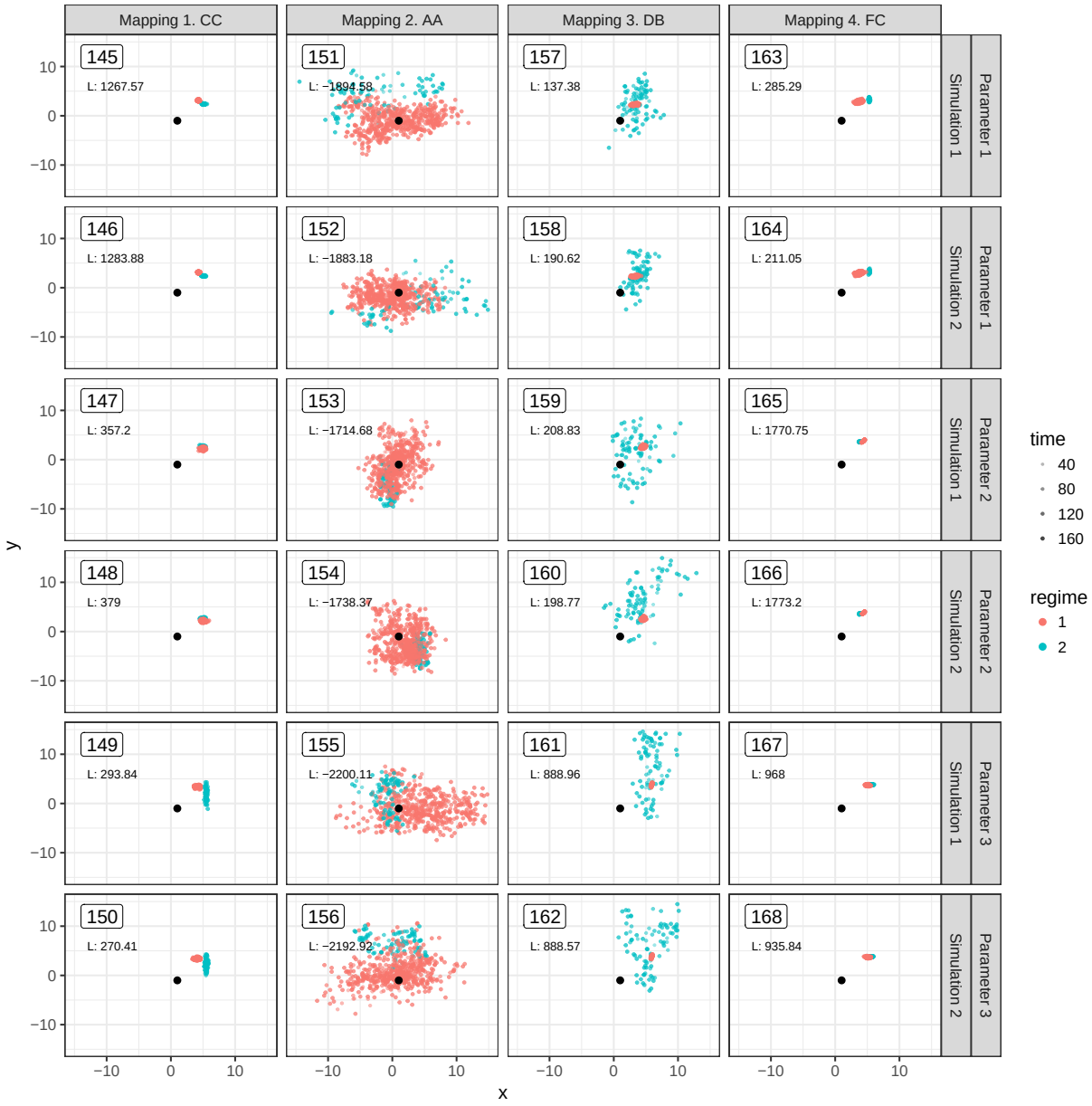


Figure S13: **Scatter plots of the simulated data-sets for big non-ultrametric trees with 2 regimes.** See legend for fig. S7.

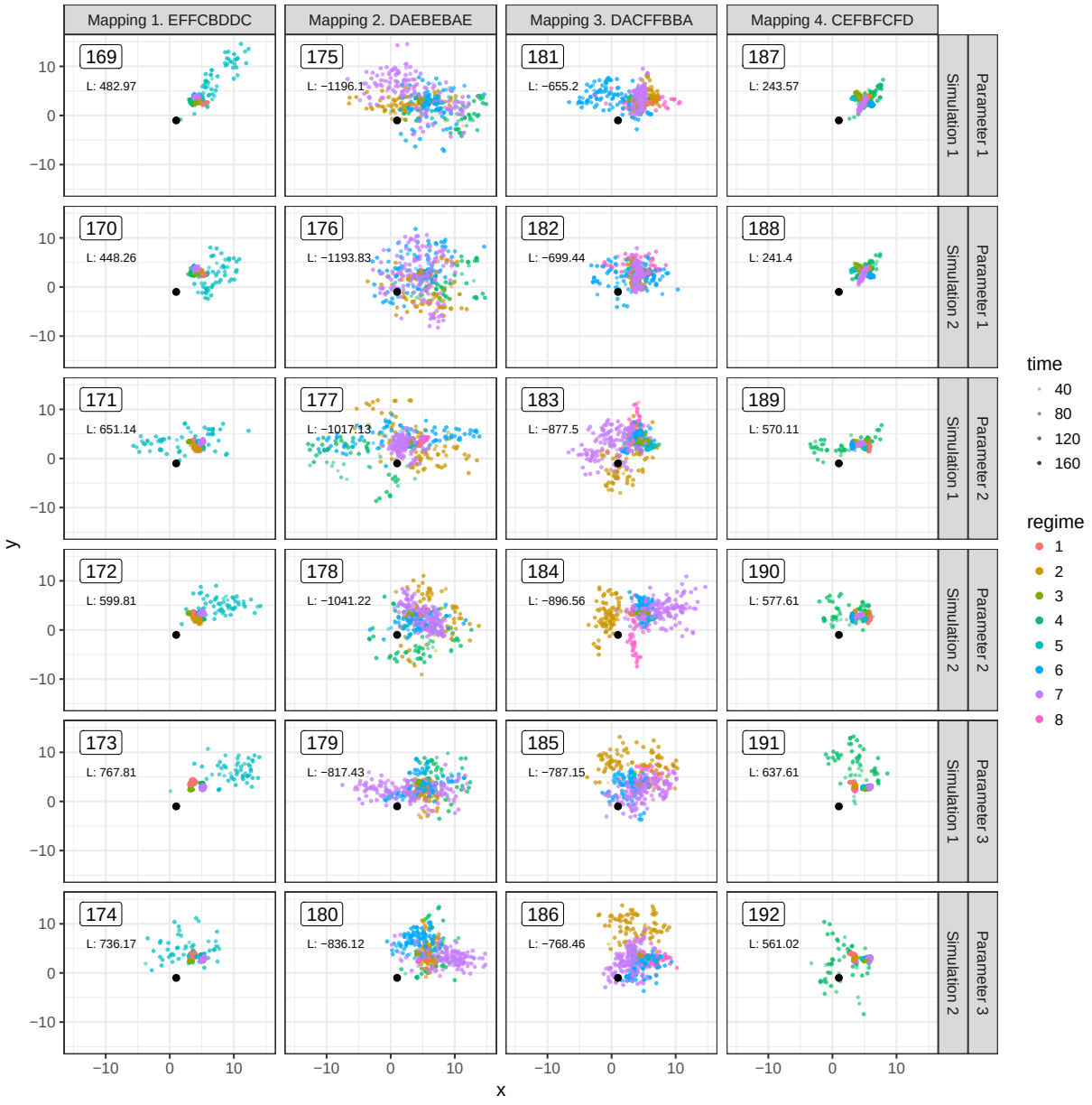


Figure S14: **Scatter plots of the simulated data-sets for big non-ultrametric trees with 8 regimes.** See legend for fig. S7.

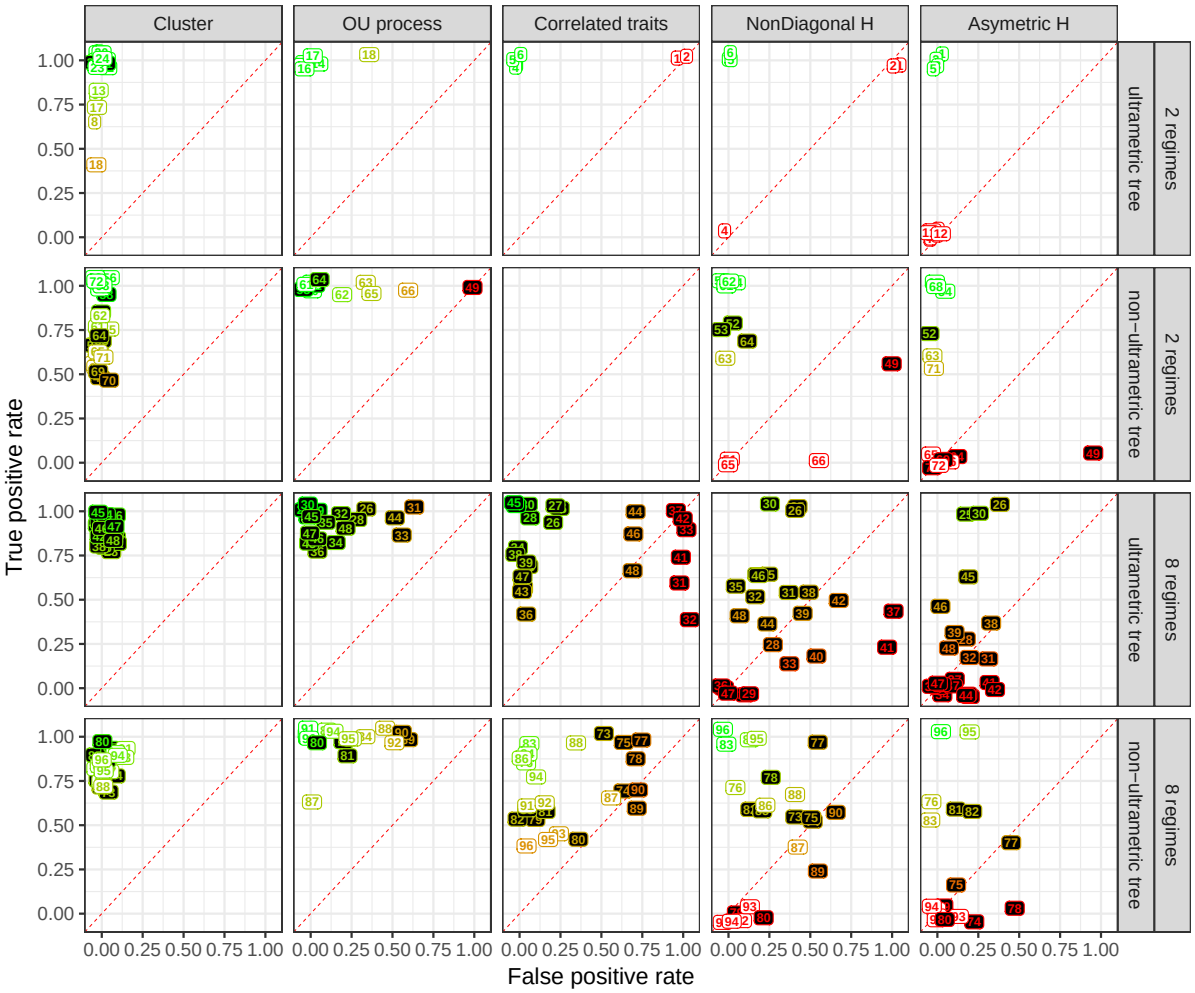


Figure S15: **Performance of the MGPM inference in simulated data for small trees.** The performance is evaluated based on the true/false positive rate for each one of the five binary criteria, explained in the text. The optimum is located at the top-left corner, i.e. when the true positive rate is equal to 1 and the false positive rate is equal to 0. The diagonal shown with a red dashed line denotes the performance to be expected from a random predictor. Numbered labels denote the result for the inferred MGPM for the corresponding simulated data (the numbers match with the ids on figs. S7-S14). Greener text denote proximity to the optimum (0, 1), whereas redder text denotes proximity to the lower right corner (1,0). The background color in each label denotes the comparison between the best inferred AIC score against the AIC score for the true model - white background denotes that the inferred model has an AIC at least as good (smaller or equal) as the AIC for the true model; black background denotes that the inferred model has a worse (bigger) AIC compared to the true model.

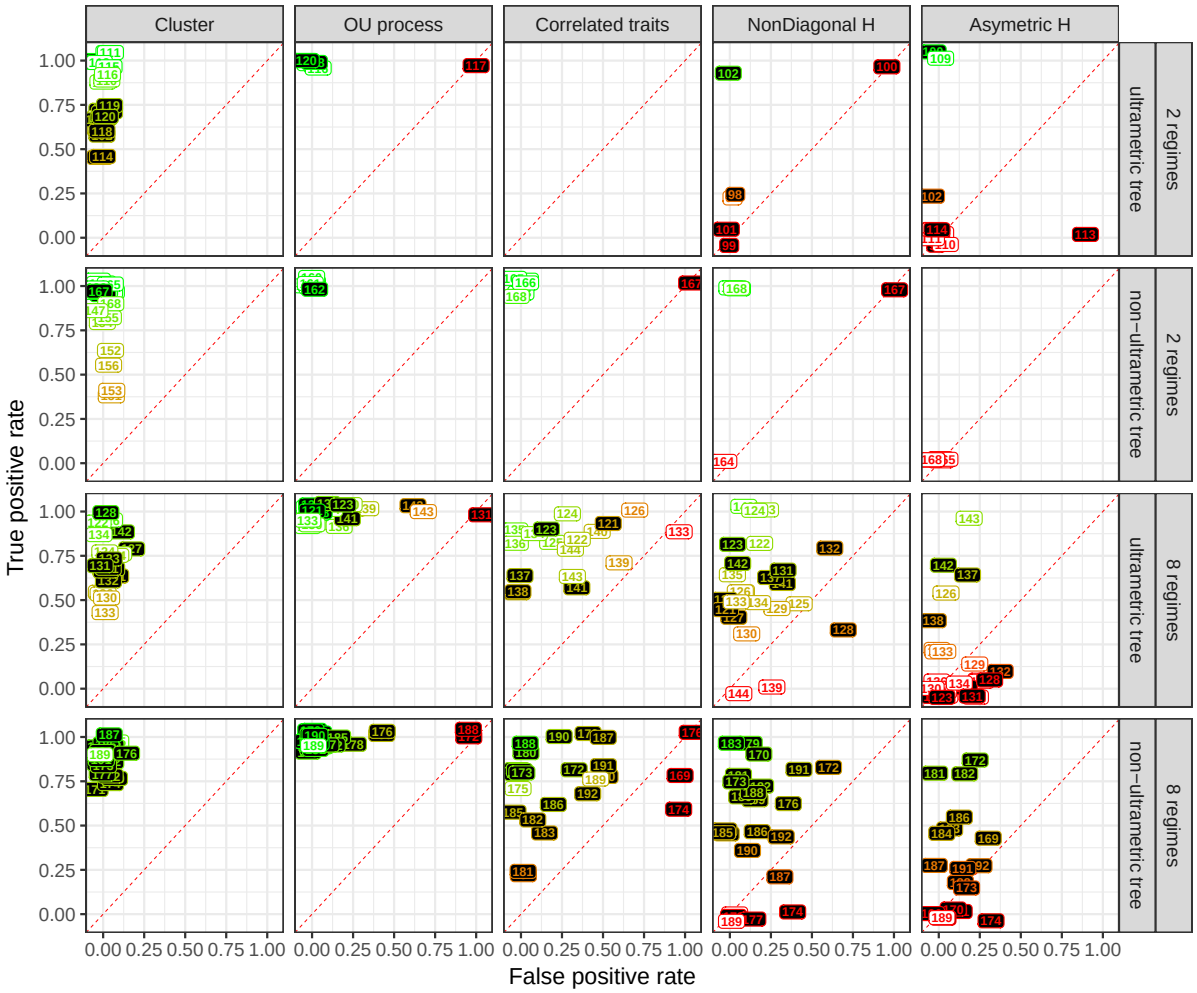


Figure S16: **Performance of the MGPM inference in simulated data for small trees.** See legend for fig. S15.

7.J SUPPLEMENTARY TABLES

Table S1: **Summary of the simulation results**

Crit.	N	#regimes	Tree-type	#tests	Better AIC	fpr	tpr
Cluster	318	2	ultrametric	24	0.92	0.00	0.92
Cluster	318	2	non-ultrametric	24	0.67	0.00	0.77
Cluster	318	8	ultrametric	24	0.00	0.03	0.87
Cluster	318	8	non-ultrametric	24	0.50	0.04	0.84
Cluster	638	2	ultrametric	24	0.42	0.00	0.78
Cluster	638	2	non-ultrametric	24	0.92	0.00	0.89
Cluster	638	8	ultrametric	24	0.58	0.02	0.74
Cluster	638	8	non-ultrametric	24	0.08	0.02	0.87
OU process	318	2	ultrametric	6	1.00	0.06	1.00
OU process	318	2	non-ultrametric	12	0.67	0.21	1.00
OU process	318	8	ultrametric	18	0.00	0.16	0.94
OU process	318	8	non-ultrametric	18	0.61	0.20	0.97
OU process	638	2	ultrametric	6	0.33	0.17	1.00
OU process	638	2	non-ultrametric	6	0.83	0.00	1.00
OU process	638	8	ultrametric	24	0.58	0.17	0.99
OU process	638	8	non-ultrametric	24	0.08	0.15	0.98
Correlated traits	318	2	ultrametric	6	1.00	0.33	1.00
Correlated traits	318	2	non-ultrametric	0			
Correlated traits	318	8	ultrametric	24	0.00	0.37	0.80
Correlated traits	318	8	non-ultrametric	24	0.50	0.29	0.71
Correlated traits	638	2	ultrametric	0			
Correlated traits	638	2	non-ultrametric	6	0.83	0.17	0.99
Correlated traits	638	8	ultrametric	18	0.67	0.30	0.80
Correlated traits	638	8	non-ultrametric	24	0.08	0.28	0.74
NonDiagonal H	318	2	ultrametric	6	1.00	0.33	0.83
NonDiagonal H	318	2	non-ultrametric	12	0.67	0.14	0.61
NonDiagonal H	318	8	ultrametric	24	0.00	0.32	0.41
NonDiagonal H	318	8	non-ultrametric	24	0.50	0.24	0.51
NonDiagonal H	638	2	ultrametric	6	0.17	0.17	0.40
NonDiagonal H	638	2	non-ultrametric	6	0.83	0.17	0.83
NonDiagonal H	638	8	ultrametric	24	0.58	0.18	0.57
NonDiagonal H	638	8	non-ultrametric	24	0.08	0.15	0.52
Asymetric H	318	2	ultrametric	12	1.00	0.00	0.33
Asymetric H	318	2	non-ultrametric	18	0.67	0.06	0.27
Asymetric H	318	8	ultrametric	24	0.00	0.15	0.23
Asymetric H	318	8	non-ultrametric	18	0.50	0.11	0.28
Asymetric H	638	2	ultrametric	12	0.42	0.07	0.18
Asymetric H	638	2	non-ultrametric	6	0.83	0.00	0.00
Asymetric H	638	8	ultrametric	24	0.58	0.11	0.17
Asymetric H	638	8	non-ultrametric	18	0.06	0.11	0.30

Part IV

POSTFACE

8

GENERAL DISCUSSION AND OUTLOOK

In this thesis I've studied in detail several challenges encountered in the transfer of classical phylogenetic comparative methods to big phylogenetically linked comparative data. Most of the tools I've developed capitalize on the idea that, for Gaussian phylogenetic models, it is possible to calculate the likelihood in linear time with respect to the number of tips in the tree, through post-order tree traversal (pruning). The idea of pruning is not novel, see e.g. (Felsenstein, 1973), but a major challenge is to generalize this idea to a large enough biologically interpretable family of models. This task is now accomplished for a large family of multiple trait models, namely $\mathcal{G}_{LInv}$, and for any type of phylogenetic tree.

In Chapter 1, I have listed the key assumptions for the parameters α and σ^2 of an Ornstein-Uhlenbeck process that need to be valid in order for the Ornstein-Uhlenbec process to be a valid model of stabilizing selection. It is unlikely that these assumptions are met in any of the contemporary applications of these models to real data (Losos, 2011). Hence, direct interpretation of parameters like α is not possible. Rather, it is needed to study the patterns in simulated data resulting from the combination of inferred model parameters and to compare these patterns with the real data used for the parameter inference. This was my approach in Chapter 3, where I've shown that the maximum likelihood parameters of an OU model fit to HIV patient data from the UK generate far more accurate patterns than a BM fit to the same data. The development of future PCMs should focus on automating this procedure of comparing observable versus simulated patterns.

In Discussion of chapter 3, I pointed out an important issue with the POUMM model, namely, the fact that the POUMM model assumes a homogeneous process of evolution along the entire tree. It is highly doubtful that such a homogeneous process could represent the dynamics in a socially heterogeneous group of HIV patients. The final result of this thesis – the mixed Gaussian phylogenetic model provides an alternative to the POUMM, which would address this issue. Hence, the application of MGPMs to the same data used for the epidemiological analysis in Chapters 3 and 4 could be of interest for future work.

With that respect, the implementation of the mixed Gaussian phylogenetic model within the PCMBase and PCMFit R-packages is an important achievement of this thesis. The presence of tools for fast inference of such a general family of models should move the focus from the technical issues of fitting complex models to big data to the conceptual issue of model identifiability and interpretability. For example, many questions regarding the identifiability of Ornstein-Uhlenbeck models are poorly understood. In particular, it is not known whether it is possible to infer jointly the trait values at the ancestral nodes and the long-term optimum in an ultrametric tree when different lineages follow different rates of evolution. Other models of evolution, such as Ornstein-Uhlenbeck models of punctuated equilibrium and acceleration/deceleration models of adaptive radiation, can easily be introduced within the PCMBase package, but need to be studied in detail before attempting their inference.

A related major challenge for future development is the quantification of the uncertainty of the model parameter estimates in fits of the mixed Gaussian phylogenetic model. Bayesian reversible jump sampling (Uyeda and Harmon, 2014) is a candidate solution, but its applicability in practice is questionnable, due to the high dimensionality of these models and the poor potential for parallelizing MCMC sampling.

Another direction of development is the application of the models described here to other types of continuous trait data. In particular, it is possible to infer MGPM models on gene expression profile comparative data. Linking patterns of evolution of the gene expression profile to other types of phenotypic data could enable discoveries of novel gene–phenotype pathways.

I will be glad to see some of the above ideas develop in real projects. My wildest hope is that some of the tools developed through this thesis would become widely used by evolutionary biologists around the world.

BIBLIOGRAPHY

Alfaro, Michael E, Francesco Santini, Chad Brock, Hugo Alamillo, Alex Dornburg, Daniel L Rabosky, Giorgio Carnevale, and Luke J Harmon (Aug. 2009). "Nine exceptional radiations plus high turnover explain species diversity in jawed vertebrates." In: *PNAS* 106.32, pp. 13410–13414.

Alizon, Samuel et al. (2010). "Phylogenetic approach reveals that virus genotype largely determines HIV set-point viral load." In: *PLoS pathogens* 6.9, e1001123.

Allaire, J J, Jonathan McPherson, Yihui Xie, Hadley Wickham, Joe Cheng, and Jeff Allen (2014). "rmarkdown: Dynamic Documents for R." In:

Anderson, Tim J C, Jeff T Williams, Shalini Nair, Daniel Sudimack, Marion Barends, Anchalee Jaidee, Ric N Price, and Francois Nosten (2010). "Inferred relatedness and heritability in malaria parasites." In: *Proceedings of the Royal Society B-Biological Sciences* 277.1693, pp. 2531–2540.

Ané, Cécile, Lam si Tung Ho, and Sebastien Roch (2016). "Phase transition on the convergence rate of parameter estimation under an Ornstein–Uhlenbeck diffusion on a tree." In: *Journal of mathematical biology* 74.1-2, pp. 1–31.

Angelino, Elaine, Eddie Kohler, Amos Waterland, Margo Seltzer, and Ryan P Adams (2014). "Accelerating MCMC via Parallel Predictive Prefetching." In: *UAI* stat.ML, arXiv:1403.7265.

Ayres, Daniel L et al. (Jan. 2012). "BEAGLE: An Application Programming Interface and High-Performance Computing Library for Statistical Phylogenetics." In: *Systematic Biology* 61.1, pp. 170–173.

Bachmann, Nadine, Teja Turk, Claus Kadelka, Alex Marzel, Mohaned Shilaih, Jürg Böni, Vincent Aubert, Thomas Klimkait, Gabriel E Leventhal, Huldrych F Günthard, et al. (May 2017). "Parent-offspring regression to estimate the heritability of an HIV-1 trait in a realistic setup." In: *Retrovirology* 14.1, p. 33.

Bartoszek, Krzysztof (Aug. 2014). "Quantifying the effects of anagenetic and cladogenetic evolution." In: *Mathematical biosciences* 254, pp. 42–57.

Bartoszek, Krzysztof and Serik Sagitov (2015). "Phylogenetic confidence intervals for the optimal trait value." In: *J. Applied Probability*.

Bartoszek, Krzysztof, Jason Pienaar, Petter Mostad, Staffan Andersson, and Thomas F Hansen (Dec. 2012). "A phylogenetic comparative method for studying multivariate adaptation." In: *Journal of theoretical biology* 314, pp. 204–215.

Bartoszek, Krzysztof, Sylvain Glémin, Ingemar Kaj, and Martin Lascoux (Sept. 2017). "Using the Ornstein-Uhlenbeck process to model the evolution of interacting populations." In: *Journal of theoretical biology* 429, pp. 35–45.

Bastide, Paul, Mahendra Mariadassou, and St 'ephane Robin (2017). "Detection of adaptive shifts on phylogenies by using shifted stochastic processes on a tree." In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 79.4, pp. 1067–1093.

Bastide, Paul, Cécile Ané, Stéphane Robin, and Mahendra Mariadassou (Jan. 2018). "Inference of Adaptive Shifts for Multivariate Correlated Traits." In: *Systematic Biology* 113.4, pp. 2158–680.

Beaulieu, Jeremy M, Dwueng-Chwuan Jhwueng, Carl Boettiger, and Brian C O'Meara (Apr. 2012). "Modeling Stabilizing Selection: Expanding The Ornstein-Uhlenbeck Model Of Adaptive Evolution." In: *Evolution; international journal of organic evolution* 66.8, pp. 2369–2383.

Bedford, Trevor and Daniel L Hartl (Jan. 2009). "Optimization of gene expression by natural selection." In: *Proceedings of the National Academy of Sciences*. Department of Organismic and Evolutionary Biology, Harvard University, 16 Divinity Avenue, Cambridge, MA 02138. National Academy of Sciences, pp. 1133–1138.

Benton, Michael J and Brent C Emerson (Jan. 2007). "How Did Life Become So Diverse? The Dynamics Of Diversification According To The Fossil Record And Molecular Phylogenetics." In: *Palaeontology* 50.1, pp. 23–40.

Bertels, Frederic, Alex Marzel, Gabriel Leventhal, Venelin Mitov, Jacques Fellay, Huldrych F Günthard, Jürg Böni, Sabine Yerly, Thomas Klimkait, Vincent Aubert, et al. (Oct. 2017). "Dissecting HIV Virulence: Heritability of Setpoint Viral Load, CD4+ T Cell Decline and Per-Parasite Pathogenicity." In: *Molecular biology and evolution*.

Bininda-Emonds, Olaf R P, Marcel Cardillo, Kate E Jones, Ross D E MacPhee, Robin M D Beck, Richard Grenyer, Samantha A Price, Rutger A Vos, John L Gittleman, and Andy Purvis (Mar. 2007). "The delayed rise of present-day mammals." In: *Nature* 446.7135, pp. 507–512.

Blanquart, François, Chris Wymant, Marion Cornelissen, Astrid Gall, Margreet Bakker, Daniela Bezemer, Matthew Hall, Mariska Hillebregt, Swee Hoe Ong, Jan Albert, et al. (June 2017). "Viral genetic variation accounts for a third of variability in HIV-1 set-point viral load in Europe." In: *Plos Biology* 15.6, e2001855.

Blomberg, Simon Phillip (Nov. 2017). "Beyond Brownian motion and the Ornstein-Uhlenbeck process: Stochastic diffusion models for the evolution of quantitative characters." In: *bioRxiv*, p. 067363.

Boddy, A M, M R McGowen, C C Sherwood, L I Grossman, M Goodman, and D E Wildman (May 2012). "Comparative analysis of encephalization in mammals reveals relaxed constraints on anthropoid primate and cetacean brain scaling." In: *Journal of Evolutionary Biology* 25.5, pp. 981–994.

Bokma, F (Oct. 2002). "Detection of punctuated equilibrium from molecular phylogenies." In: *Journal of Evolutionary Biology* 15.6, pp. 1048–1056.

Bortolussi, Nicolas, Eric Durand, Michael Blum, and Olivier Francois (2012). "apTreeshape: Analyses of Phylogenetic Treeshape." In: *R package*.

Boskova, Veronika, Sebastian Bonhoeffer, and Tanja Stadler (2014). "Inference of Epidemiological Dynamics Based on Simulated Phylogenies Using Birth-Death and Coalescent Models." In: *PLoS Computational Biology (PLOSCB) 10(4)* 10.11, e1003913.

Boucher, Florian C, Vincent Démery, Elena Conti, Luke J Harmon, and Josef Uyeda (Mar. 2018). "A General Model for Estimating Macroevolutionary Landscapes." In: *Systematic Biology* 67.2, pp. 304–319.

Bouckaert, Remco R, Joseph Heled, Denise Kühnert, Timothy G Vaughan, Chieh-Hsi Wu, Dong Xie, Marc A Suchard, Andrew Rambaut, and Alexei J Drummond (2014). "BEAST 2 - A Software Platform for Bayesian Evolutionary Analysis." In: *PLoS Computational Biology (PLOSCB) 10(4)* 10.4, e1003537–.

Boyd, Stephen P and Lieven Vandenberghe (Mar. 2004). *Convex Optimization*. Cambridge University Press.

Brockwell, A E (Mar. 2006). "Parallel Markov chain Monte Carlo simulation by pre-fetching." In: *Journal of Computational and Graphical Statistics* 15.1, pp. 246–261.

Brooks, S P and A Gelman (Dec. 1998). "General methods for monitoring convergence of iterative simulations." In: *Journal of Computational and Graphical Statistics* 7.4, pp. 434–455.

Butler, M A and A A King (Dec. 2004). "Phylogenetic comparative analysis: A modeling approach for adaptive evolution." In: *American Naturalist* 164.6, pp. 683–695.

Byrd, Richard H, Peihuang Lu, Jorge Nocedal, and Ci You Zhu (1995). "A limited memory algorithm for bound constrained optimization." In: *SIAM Journal on Scientific Computing* 16.5, pp. 1190–1208.

Clavel, Julien, Gilles Escarguel, and Gildas Merceron (Nov. 2015). "mvmorph: an r package for fitting multivariate evolutionary models to morphometric data." In: *Methods in Ecology and Evolution* 6.11, pp. 1311–1319.

Cook, Samantha R, Andrew Gelman, and Donald B Rubin (Sept. 2006). "Validation of Software for Bayesian Models Using Posterior Quantiles." In: *Journal of Computational and Graphical Statistics* 15.3, pp. 675–692.

Cooper, Natalie, Gavin H Thomas, Chris Venditti, Andrew Meade, and Rob P Freckleton (Dec. 2015). "A cautionary note on the use of Ornstein Uhlenbeck models in macroevolutionary studies." In: *Biological Journal of the Linnean Society* 118.1, pp. 64–77.

Cressler, Clayton E, Marguerite A Butler, and Aaron A King (Nov. 2015). "Detecting Adaptive Evolution in Phylogenetic Comparative Analysis Using the Ornstein-Uhlenbeck Model." In: *Systematic Biology* 64.6, pp. 953–968.

Dahl, David B. "xtable: Export Tables to LaTeX or HTML." In: ().

Dowle, Matthew, T Short, S Liangolou, and A Srinivasan (July 2014). "data.table: Extension of data.frame." In: *unpubli*, p. 9.

Drummond, Alexei J, Marc A Suchard, Dong Xie, and Andrew Rambaut (Aug. 2012). "Bayesian phylogenetics with BEAUti and the BEAST 1.7." In: *Molecular biology and evolution* 29.8, pp. 1969–1973.

Drury, Jonathan, Julien Clavel, Marc Manceau, and Helene Morlon (July 2016). "Estimating the Effect of Competition on Trait Evolution Using Maximum Likelihood Inference." In: *Systematic Biology* 65.4, pp. 700–710.

Duchen, Pablo, Christoph Leuenberger, Sándor M Szilágyi, Luke Harmon, Jonathan Eastman, Manuel Schweizer, and Daniel Wegmann (Nov. 2017). "Inference of Evolutionary Jumps in Large Phylogenies using Lévy Processes." In: *Systematic Biology* 66.6, pp. 950–963.

Eastman, Jonathan M, Michael E Alfaro, Paul Joyce, Andrew L Hipp, and Luke J Harmon (Dec. 2011). "A Novel Comparative Method For Identifying Shifts In The Rate Of Character Evolution On Trees." In: *Evolution* 65.12, pp. 3578–3589.

Eddelbuettel, Dirk. "digest: Create Compact Hash Digests of R Objects." In: ().

— (June 2013). *Seamless R and C++ Integration with Rcpp*. New York, NY: Springer Science & Business Media.

Eddelbuettel, Dirk and Conrad Sanderson (2014). "RcppArmadillo - Accelerating R with high-performance C++ linear algebra." In: *Computational Statistics & Data Analysis* 71, pp. 1054–1063.

Edwards, A W F (1970). "Estimation of the branch points of a branching diffusion process. (With discussion.)" In: *Journal of the Royal Statistical Society. Series B. Methodological* 32, pp. 155–174.

Eldredge, N and S J Gould (1972). "Punctuated equilibria: an alternative to phyletic gradualism." In: *Models in paleobiology*. Ed. by T J M Schopf and J M Thomas. San Francisco: Freeman Cooper, pp. 82–115.

Falconer, D S (Feb. 1996). *Introduction to Quantitive Genetics*. San Val, Incorporated.

Felsenstein, J (Sept. 1973). "Maximum-likelihood estimation of evolutionary trees from continuous characters." In: *American Journal of Human Genetics* 25.5, pp. 471–492.

— (1981). "Evolutionary trees from DNA sequences: a maximum likelihood approach." In: *Journal of molecular evolution* 17.6, pp. 368–376.
— (Jan. 1988). "Phylogenies And Quantitative Characters." In: *Annual Review of Ecology and Systematics* 19.1, pp. 445–471.
Felsenstein, Joseph (1983). "Statistical Inference of Phylogenies." In: *Journal of the Royal Statistical Society. Series A (General)* 146.3, p. 246.
— (Jan. 1985). "Phylogenies and the Comparative Method." In: *The American Naturalist* 125.1, pp. 1–15.
FitzJohn, Richard G (Dec. 2010). "Quantitative traits and diversification." In: *Systematic Biology* 59.6, pp. 619–633.
— (Dec. 2012). "Diversitree: comparative phylogenetic analyses of diversification in R." In: *Methods in Ecology and Evolution* 3.6, pp. 1084–1092.
Freckleton, Robert P (July 2012). "Fast likelihood calculations for comparative analyses." In: *Methods in Ecology and Evolution* 3.5, pp. 940–947.
Gavryushkina, Alexandra, David Welch, Tanja Stadler, and Alexei J Drummond (Dec. 2014). "Bayesian Inference of Sampled Ancestor Trees for Epidemiology and Fossil Calibration." In: *PLoS Computational Biology (PLOSCB) 10(4)* 10.12.
Genz, Alan and Frank Bretz (July 2009). *Computation of Multivariate Normal and t Probabilities*. Springer Science & Business Media.
Golub, G H and C F Van Loan (2013). *Matrix Computations*. Baltimore: The Johns Hopkins University Press.
Golub, Gene H and Charles F Van Loan (Dec. 2012). *Matrix Computations*. JHU Press.
Goolsby, Eric W, Jorn Bruggeman, and Cécile Ané (July 2016). "Rphylopars: fast multivariate phylogenetic comparative methods for missing data and within-species variation." In: *Methods in Ecology and Evolution* 8.1, pp. 22–27.
Goudie, Robert J B, Rebecca M Turner, Daniela De Angelis, and Andrew Thomas (Apr. 2017). "MultiBUGS: Massively parallel MCMC for Bayesian hierarchical models." In: *arXiv.org*, arXiv:1704.03216. arXiv: 1704.03216 [stat.CO].
Gould, S J and N Eldredge (Nov. 1993). "Punctuated equilibrium comes of age." In: *Nature* 366.6452, pp. 223–227.
Goulet, Vincent, Christophe Dutang, Martin Maechler, David Firth, Marina Shapir, and Michael Stadelmann. "expm: Matrix Exponential, Log." In: ().
Gouri 'eroux, C, A Monfort, and A Trognon (1984). "Pseudomaximum likelihood methods: theory." In: *Econometrica. Journal of the Econometric Society* 52.3, pp. 681–700.
Grimmett, Geoffrey and David Stirzaker (May 2001). *Probability and Random Processes*. Oxford University Press.
Haario, Heikki, Eero Saksman, and Johanna Tamminen (2001). "An adaptive metropolis algorithm." In: *Bernoulli. Official Journal of the Bernoulli Society for Mathematical Statistics and Probability* 7.2, pp. 223–242.
Hansen, Thomas F (Oct. 1997). "Stabilizing Selection and the Comparative Analysis of Adaptation." In: *Evolution; international journal of organic evolution* 51.5, pp. 1341–1351.
Hansen, Thomas F and Krzysztof Bartoszek (May 2012). "Interpreting the evolutionary regression: the interplay between observational and biological errors in phylogenetic comparative studies." In: *Systematic Biology* 61.3, pp. 413–425.
Hansen, Thomas F and Emília P Martins (Aug. 1996). "Translating Between Microevolutionary Process and Macroevolutionary Patterns: The Correlation Structure of Interspecific Data." In: *Evolution* 50.4, p. 1404.
Hansen, Thomas F, Jason Pienaar, and Steven Hecht Orzack (Aug. 2008). "A comparative method for studying adaptation to a randomly evolving environment." In: *Evolution* 62.8, pp. 1965–1977.
Harmon, Luke J (2018). *Phylogenetic Comparative Methods*. Learning from trees.
Harmon, Luke J, Jason T Weir, Chad D Brock, Richard E Glor, and Wendell Challenger (Jan. 2008). "GEIGER: investigating evolutionary radiations." In: *Bioinformatics* 24.1, pp. 129–131.
Hartl, D L and A G Clark (2007). *Principles of population genetics*. Sinauer Associates.
Ho, Lam si Tung and Cécile Ané (Apr. 2013). "Asymptotic theory with hierarchical autocorrelation: Ornstein–Uhlenbeck tree models." In: *The Annals of Statistics* 41.2, pp. 957–981.
— (Apr. 2014a). "A linear-time algorithm for Gaussian and non-Gaussian trait evolution models." In: *Systematic Biology* 63.3, pp. 397–408.
— (Nov. 2014b). "Intrinsic inference difficulties for trait evolution with Ornstein-Uhlenbeck models." In: *Methods in Ecology and Evolution* 5.11, pp. 1133–1146.
Hodcroft, Emma, Jarrod D Hadfield, Esther Fearnhill, Andrew Phillips, David Dunn, Siobhan O'Shea, Deenan Pillay, and Andrew J Leigh Brown (May 2014). "The Contribution of Viral Genotype to Plasma Viral Set-Point in HIV Infection." In: *PLoS pathogens* 10.5, e1004112.
Housworth, Elizabeth A, Emília P Martins, and Michael Lynch (Jan. 2004). "The phylogenetic mixed model." In: *The American Naturalist* 163.1, pp. 84–96.
Ingram, Travis and D Luke Mahler (May 2013). "SURFACE: detecting convergent evolution from comparative data by fitting Ornstein-Uhlenbeck models with stepwise Akaike Information Criterion." In: *Methods in Ecology and Evolution* 4.5, pp. 416–425.

Ives, Anthony R and Theodore Jr Garland (Jan. 2010). "Phylogenetic Logistic Regression for Binary Dependent Variables." In: *Systematic Biology* 59.1, pp. 9–26.

Jacquard, A (June 1983). "Heritability: One Word, Three Concepts." In: *Biometrics* 39.2, p. 465.

Jerison, Harry (1973). *Evolution of The Brain and Intelligence*. New York: Academic press, Inc.

Jhwueng, Dwueng-Chwuan and Brian OMeara (2015). "Trait Evolution on Phylogenetic Networks." In: *bioRxiv*, p. 023986.

Khabbazian, Mohammad, Ricardo Kriebel, Karl Rohe, and Cécile Ané (Feb. 2016). "Fast and accurate detection of evolutionary shifts in Ornstein-Uhlenbeck models." In: *Methods in Ecology and Evolution* 7.7, pp. 811–824.

Kühnert, Denise, Tanja Stadler, Timothy G Vaughan, and Alexei J Drummond (Apr. 2016). "Phylodynamics with Migration: A Computational Framework to Quantify Population Structure from Genomic Data." In: *Molecular biology and evolution* 33.8, msw064–2116.

Kutsukake, Nobuyuki and Hideki Innan (Feb. 2013). "Simulation-Based Likelihood Approach for Evolutionary Models of Phenotypic Traits on Phylogeny." In: *Evolution; international journal of organic evolution* 67.2, pp. 355–367.

Lande, R (1976). "Natural-Selection and Random Genetic Drift in Phenotypic Evolution." In: *Evolution; international journal of organic evolution* 30.2, pp. 314–334.

Landis, Michael J, Joshua G Schraiber, and Mason Liang (Dec. 2012). "Phylogenetic Analysis Using Lévy Processes: Finding Jumps in the Evolution of Continuous Traits." In: *Systematic Biology* 62.2, pp. 193–204.

Le Gall, F (2014). "Powers of tensors and fast matrix multiplication." In: *Proceedings of the 39th International Symposium on Symbolic and Algebraic Computation, Kobe, Japan*, pp. 296–303.

Losos, Jonathan B (June 2011). "Seeing the forest for the trees: the limitations of phylogenies in comparative biology. (American Society of Naturalists Address)." In: *The American Naturalist* 177.6, pp. 709–727.

Lynch, Michael (Aug. 1991). "Methods for the Analysis of Comparative Data in Evolutionary Biology." In: *Evolution; international journal of organic evolution* 45.5, pp. 1065–1080.

Lynch, Michael and Bruce Walsh (Jan. 1998). *Genetics and Analysis of Quantitative Traits*. Sinauer Associates Incorporated.

Manceau, Marc, Amaury Lambert, and Helene Morlon (Dec. 2016). "A unifying comparative phylogenetic framework including traits coevolving across interacting lineages." In: *Systematic Biology* 66.4, syw115–568.

Martins, Emília P and Thomas F Hansen (Apr. 1997). "Phylogenies and the Comparative Method: A General Approach to Incorporating Phylogenetic Information into the Analysis of Interspecific Data." In: *The American Naturalist* 149.4, p. 646.

Mayr, Ernst (Nov. 1982). "SPECIATION AND MACROEVOLUTION." In: *Evolution* 36.6, pp. 1119–1132.

Metropolis, Nicholas, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller (June 1953). "Equation of State Calculations by Fast Computing Machines." In: *The Journal of Chemical Physics* 21.6, pp. 1087–1092.

Microsoft and Steve Weston. "foreach: Foreach Looping Construct for R." In: ().

Mitov, Venelin, Krzysztof Bartoszek, and Tanja Stadler. "Automatic Generation of Evolutionary Hypotheses using Mixed Gaussian Phylogenetic Models." In: *preprint* ().

Mitov, Venelin and Tanja Stadler (June 2016). "The heritability of pathogen traits - definitions and estimators." In: *preprint* `https://www.biorxiv.org/content/early/2016/06/12/058503`. Last accessed June 12, 2016.

— (2017a). "Fast Bayesian Inference of Phylogenetic Models Using Parallel Likelihood Calculation and Adaptive Metropolis Sampling." In: *preprint*, p. 235739. eprint: `10.1101/235739`.

— (2017b). "POUMM: An R-package for Bayesian Inference of Phylogenetic Heritability." In:

— (Jan. 2018). "A Practical Guide to Estimating the Heritability of Pathogen Traits." In: *Molecular biology and evolution* 6.9, e1001123.–msx328 VL –IS –.

Mitov, Venelin, Krzysztof Bartoszek, Georgios Asimomitis, and Tanja Stadler (2018). "Fast Likelihood Calculation For Multivariate Phylogenetic Comparative Methods: The PCMBase R Package." In: *preprint*.

Montgomery, Stephen H, Isabella Capellini, Robert A Barton, and Nicholas I Mundy (Jan. 2010). "Reconstructing the ups and downs of primate brain evolution: implications for adaptive hypotheses and Homo floresiensis." In: *BMC biology* 8.1, p. 9.

Müller, Nicola F, David A Rasmussen, and Tanja Stadler (June 2017). "The Structured Coalescent and Its Approximations." In: *Molecular biology and evolution* 34.11, pp. 2970–2981.

Nuismer, Scott L and Luke J Harmon (Jan. 2015). "Predicting rates of interspecific interaction from phylogenetic trees." In: *Ecology letters* 18.1, pp. 17–27.

O'Meara, Brian C (Nov. 2012). "Evolutionary Inferences from Phylogenies: A Review of Methods." In: *Annual Review of Ecology, Evolution, and Systematics* 43.1, pp. 267–285.

O'Meara, Brian C, Cécile Ané, Michael J Sanderson, and Peter C Wainwright (May 2006). "Testing for different rates of continuous trait evolution using likelihood." In: *Evolution* 60.5, pp. 922–933.

Ornstein, L S and F Zernike (1919). "The theory of the Brownian Motion and statistical mechanics." In: *Proceedings of the Koninklijke Akademie Van Wetenschappen Te Amsterdam* 21.1/5, pp. 109–114.

Pagel, M (1994). "Detecting Correlated Evolution on Phylogenies - a General-Method for the Comparative-Analysis of Discrete Characters." In: *Proceedings of the Royal Society B-Biological Sciences* 255.1342, pp. 37–45.

Paradis, E, J Claude, and K Strimmer (2004). "APE: Analyses of Phylogenetics and Evolution in R language." In: *Bioinformatics* 20.2, pp. 289–290.

Paradis, Emmanuel and Julien Claude (2002). "Analysis of comparative data using generalized estimating equations." In: *Journal of theoretical biology* 218.2, pp. 175–185.

Pennell, Matthew W and Luke J Harmon (June 2013). "An integrative view of phylogenetic comparative methods: connections to population genetics, community ecology, and paleobiology." In: *Annals of the New York Academy of Sciences* 1289.1, pp. 90–105.

Pennell, Matthew W, Jonathan M Eastman, Graham J Slater, Joseph W Brown, Josef C Uyeda, Richard G FitzJohn, Michael E Alfaro, and Luke J Harmon (Aug. 2014). "geiger v2.0: an expanded suite of methods for fitting macroevolutionary models to phylogenetic trees." In: *Bioinformatics* 30.15, pp. 2216–2218.

Plummer, Martyn, Nicky Best, Kate Cowles, and Karen Vines (2006). "CODA: Convergence Diagnosis and Output Analysis for MCMC." In: *R News* 6, pp. 7–11.

Price, Morgan N, Paramvir S Dehal, and Adam P Arkin (July 2009). "FastTree: computing large minimum evolution trees with profiles instead of a distance matrix." In: *Molecular biology and evolution* 26.7, pp. 1641–1650.

Pybus, Oliver G et al. (Sept. 2012). "Unifying the spatial epidemiology and molecular evolution of emerging epidemics." In: *PNAS* 109.37, pp. 15066–15071.

Qamnieh, Manar (Jan. 2015). "Scheduling of Parallel Real-time DAG Tasks on Multiprocessor Systems." PhD thesis. igm.univ-mlv.fr.

Reif, John H (1989). *Parallel Algorithms Derivation*. Tech. rep. Fort Belvoir, VA: US Dept of the Navy, Funding.

Reitan, Trond, Tore Schweder, and Jorijntje Henderiks (2012). "Phenotypic evolution studied by layered stochastic differential equations." In: *The Annals of Applied Statistics* 6.4, pp. 1531–1551.

Revell, Liam J (Dec. 2011). "phytools: an R package for phylogenetic comparative biology (and other things)." In: *Methods in Ecology and Evolution* 3.2, pp. 217–223.

Revolution Analytics and Weston, Steve. "iterators: Iterator Construct for R." In: ().

Rohlfs, Rori V, Patrick Harrigan, and Rasmus Nielsen (Oct. 2013). "Modeling Gene Expression Evolution with an Extended Ornstein–Uhlenbeck Process Accounting for Within-Species Variation." In: *Molecular biology and evolution* 31.1, pp. 201–211.

Ronquist, F and J P Huelsenbeck (Aug. 2003). "MrBayes 3: Bayesian phylogenetic inference under mixed models." In: *Bioinformatics* 19.12, pp. 1572–1574.

Sanderson, Conrad and Ryan Curtin (June 2016). "Armadillo: a template-based C++ library for linear algebra." In: *Journal of Open Source Software* 1.2.

Scheidegger, Andreas (June 2012). "adaptMCMC v1.1." In: *R package*.

Shirreff, George, Samuel Alizon, Anne Cori, Huldrych F Günthard, Oliver Laeyendecker, Ard van Sighem, Daniela Bezemer, and Christophe Fraser (Sept. 2013). "How effectively can HIV phylogenies be used to measure heritability?" In: *Evolution, Medicine, and Public Health* 2013.1, pp. 209–224.

Simpson, George Gaylord (1953). *The Major Features of Evolution*.

Slater, Graham J (Aug. 2013). "Phylogenetic evidence for a shift in the mode of mammalian body size evolution at the Cretaceous-Palaeogene boundary." In: *Methods in Ecology and Evolution* 4.8, pp. 734–744.

— (May 2014). "Correction to 'Phylogenetic evidence for a shift in the mode of mammalian body size evolution at the Cretaceous-Palaeogene boundary', and a note on fitting macroevolutionary models to comparative paleontological data sets." In: *Methods in Ecology and Evolution* 5.7, pp. 714–718.

Slater, Graham J, Luke J Harmon, and Michael E Alfaro (Dec. 2012). "Integrating Fossils With Molecular Phylogenies Improves Inference Of Trait Evolution." In: *Evolution; international journal of organic evolution* 66.12, pp. 3931–3944.

Slater, Graham J, Luke J Harmon, Daniel Wegmann, Paul Joyce, Liam J Revell, and Michael E Alfaro (Mar. 2012). "Fitting Models of Continuous Trait Evolution to Incompletely Sampled Comparative Data Using Approximate Bayesian Computation." In: *Evolution; international journal of organic evolution* 66.3, pp. 752–762.

Snell, Otto (1891). "Das Gewicht des Gehirnes und des Hirnmantels der Säugerthiere in Beziehung zu deren geistigen Fähigkeiten." In: *Sitzungsberichte der Gesellschaft für Morphologie und Psychologie*, pp. 1–5.

Stadler, Tanja (Nov. 2009). "On incomplete sampling under birth-death models and connections to the sampling-based coalescent." In: *Journal of theoretical biology* 261.1, pp. 58–66.

— (Oct. 2011). "Simulating Trees with a Fixed Number of Extant Species." In: *Systematic Biology* 60.5, pp. 676–684.

Stadler, Tanja, Denise Kühnert, Sebastian Bonhoeffer, and Alexei J Drummond (Jan. 2013). "Birth-death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV)." In: *PNAS* 110.1, pp. 228–233.

Stamatakis, Alexandros, Paul Hoover, and Jacques Rougemont (Oct. 2008). "A rapid bootstrap algorithm for the RAxML Web servers." In: *Systematic Biology* 57.5, pp. 758–771.

Stebbins, G L and F J Ayala (Aug. 1981). "Is a new evolutionary synthesis necessary?" In: *Science (New York, N.Y.)* 213.4511, pp. 967–971.

Uhlenbeck, G E and L S Ornstein (Sept. 1930). "On the Theory of the Brownian Motion." In: *Physical Review* 36.5, pp. 823–841.

Uyeda, Josef C and Luke J Harmon (Nov. 2014). "A Novel Bayesian Method for Inferring and Interpreting the Dynamics of Adaptive Landscapes from Phylogenetic Comparative Data." In: *Systematic Biology* 63.6, pp. 902–918.

Vihola, Matti (Sept. 2012). "Robust adaptive Metropolis algorithm with coerced acceptance rate." In: *Statistics and Computing* 22.5, pp. 997–1008.

Wang, Qian, Xianyi Zhang, Yunquan Zhang, and Qing Yi (2013). "AUGEM - automatically generate high performance dense linear algebra kernels on x86 CPUs." In: *SC*, pp. 1–12.

Weston, Steve. "doMPI: Foreach Parallel Adaptor for the Rmpi Package." In: ().

Whiley, Matt and Simon P Wilson (2004). "Parallel algorithms for Markov chain Monte Carlo methods in latent spatial Gaussian models." In: *Statistics and Computing* 14.3, pp. 171–179.

Wickham, Hadley (2009). "ggplot2 - Elegant Graphics for Data Analysis." In: *Use R*.

Wilke, Claus O. "cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2' [R package cowplot version 0.9.3]." In: ().

Wright, S (Mar. 1931). "Evolution in Mendelian populations." In: *Genetics* 16.2, pp. 0097–0159.

Xie, Yihui (July 2017). *Dynamic Documents with R and knitr, Second Edition*. CRC Press.

Yu, Guangchuang. "ggimage: Use Image in 'ggplot2'." In: ().

Yu, Guangchuang, David K Smith, Huachen Zhu, Yi Guan, and Tommy Tsan-Yuk Lam (Jan. 2017). "GGTREE: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data." In: *Methods in Ecology and Evolution* 8.1, pp. 28–36.

Ziegler, Andreas (June 2011). *Generalized Estimating Equations*. Vol. 204. Lecture Notes in Statistics. New York, NY: Springer Science & Business Media.

Zwiernik, Piotr, Caroline Uhler, and Donald Richards (Aug. 2014). "Maximum Likelihood Estimation for Linear Gaussian Covariance Models." In: *arXiv.org*. arXiv: `1408.5604v2 [math.ST]`.

COLOPHON

This document was typeset using the typographical look-and-feel `classicthesis` developed by André Miede. The style was inspired by Robert Bringhurst's seminal book on typography "*The Elements of Typographic Style*".