Noname manuscript No.
(will be inserted by the editor)

The species problem from the modeler's point of view

Marc Manceau · Amaury Lambert

the date of receipt and acceptance should be inserted later

Keywords Genealogy · Phylogeny · Microevolution · Macroevolution · Individual-based model · Species concept

Abstract How to define a partition of individuals into species is a long-standing question called the *species problem* in systematics. Here, we focus on this problem in the thought experiment where individuals reproduce clonally and both the differentiation process and the population genealogies are explicitly known. We specify three desirable properties of species partitions: (A) Heterotypy between species, (B) Homotypy within species and (M) Monophyly of each species. We then ask: How and when is it possible to delineate species in a way satisfying these properties?

We point out that the three desirable properties cannot in general be satisfied simultaneously, but that any two of them can. We mathematically prove the existence of the finest partition satisfying (A) and (M) and the coarsest partition satisfying (B) and (M). For each of them, we propose a simple algorithm to build the associated phylogeny out of the genealogy.

Marc Manceau · Amaury Lambert

Center for Interdisciplinary Research in Biology (CIRB), Collège de France, PSL Research University, CNRS UMR 7241, INSERM U1050, 75005 Paris, France

Marc Manceau

Institut de Biologie de l'École Normale Supérieure, École Normale Supérieure, PSL Research University, CNRS UMR 8197, INSERM U1024, 75005 Paris, France

Amaury Lambert

Laboratoire de Probabilités, Statistique et Modélisation (LPSM), Sorbonne Université, CNRS UMR 8001, 75005 Paris, France

The ways we propose to phrase the species problem shed new light on the interaction between the genealogical and phylogenetic scales in modeling work. The two definitions centered on the monophyly property can readily be used at a higher taxonomic level as well, e.g. to cluster species into monophyletic genera.

Corresponding author

Marc Manceau, E-mail: marc.manceau@ens.fr, Phone: (+33)6 37 48 58 57

Orcid numbers

Marc Manceau : 0000-0001-7368-718X

Amaury Lambert : 0000-0002-7248-9955

Acknowledgements The authors are very grateful to F. Débarre, R.S. Etienne, M. Steel, S. Türpitz and A. Hoppe for their comments on this paper, and to D. Baum for helpful literature advice. The authors thank the Center for Interdisciplinary Research in Biology (CIRB, Collège de France) for funding, as well as the École Normale Supérieure for MM PhD funding. We declare no conflict of interest.

Introduction

Models in macro-evolution have traditionally been centered on species. The so-called *lineage-based models* of diversification form a wide class of models considering species as key evolutionary units, thought of as particles that can give birth to other particles (i.e., speciate) during a given lifetime (i.e., before extinction) (see Stadler 2013; Pyron and Burbrink 2013; Morlon 2014 for reviews). In contrast, evolutionary processes amenable to direct empirical measurement (differentiation, reproduction, selection) are usually described at the level of individuals or populations.

The Neutral Theory of Biodiversity (NTB) (Hubbell 2001) opened a new way of thinking about species in macro-evolution. The birth, death, differentiation and speciation processes are described at the level of individuals, under the assumption of selective neutrality. In the last two decades, a popular way of studying macro-evolution has followed, consisting in performing computer-intensive simulations of individual-based stochastic processes of species diversification (Jabot and Chave 2009; Aguilée et al. 2011; Rosindell et al. 2015; Gascuel et al. 2015; Missa et al. 2016). These

models rely on three major steps: (i) The genealogy of individuals is produced under a stochastic scenario of population dynamics; (ii) A process of phenotypic differentiation superimposed on the genealogy generates a partition of individuals into phenotypic groups; (iii) A species definition is postulated and is used to cluster individuals into different species, in relation to both the genealogy and phenotypic groups. These three steps allow modelers to track the evolutionary history of species, where extinction and speciation events emerge from the genealogical history of individual organisms.

The scenario of population dynamics (i) has for example been modeled with the Wright-Fisher or the Moran model from population genetics or with density-dependent branching processes (Durrett 2008). We will not focus on that step, but we will nonetheless consider throughout the paper that individuals reproduce clonally. The genealogical relationships within a sample $\mathcal{X}$ of present-day individuals will thus be a known rooted tree denoted T. Each tip of T is labelled by an element of $\mathcal{X}$, each internal vertex corresponds to some ancestor of elements of $\mathcal{X}$, and an edge between vertices represents a parent-child relationship. After running through step (ii), all individuals in $\mathcal{X}$ can be grouped into clusters of individuals on the basis of their phenotype. We call the associated $\mathcal{X}$-partition the *phenotypic partition*, denoted $\mathscr{P}$.

In this paper, we review how steps (ii) and (iii) have been handled in the literature, before asking the following theoretical question:

> *How* and *when* is it possible to delineate species, in a way satisfying *biologically meaningful properties*, in an ideal situation where the phenotypic partition $\mathscr{P}$ is specified and the entire genealogy T is known?

Because we work under the simplifying assumption of clonally reproducing organisms, this question is formally identical to the problem of defining and delineating *genera* when the *species phylogeny* is known, or any similar question formulated at a higher-order level (Aldous et al. 2008, 2011).

In the biological literature, the problem of agreeing on what should be considered as the most relevant concept of species is a long-standing question called the *species problem*. The species problem is both a conceptual question (defining the species concept) and a practical problem (classifying individuals into species) (Bock 2004). Several of the most notable evolutionary biologists (e.g. Dar-

win, Dobzhansky, Mayr, Simpson, Hennig...) took a stand on the species problem often leading to vigorous debates (see Mayden 1997; De Queiroz 2007 for overviews of historical disagreements).

In the very simple setting that we are looking at, two classes of species concepts appear relevant in the quest for *biologically meaningful properties* of species. The 'typological species concepts' (Regan 1925; Sneath 1976) correspond to the clustering of individuals on the sole basis of their observed phenotype. This can be translated into two desirable properties: (A) any two individuals in distinct clusters differ for at least one characteristic, (B) individuals belonging to the same cluster all share the same characteristics. The later foundation and spread of cladistics by Hennig (Hennig 1965) marked a radical change of paradigm in the systematic classification, which quickly resulted in the proposition of so-called 'phylogenetic species concepts' (De Queiroz and Donoghue 1988; Avise and Ball 1990; Baum 2009). These definitions, which brought to the forefront the notion of common ancestry, provide us with a third desirable property: (M) species are monophyletic groups of individuals, i.e. any two individuals in one cluster are more closely related to each other than either is to any individual in another cluster.

Ideally, defining putative species in our framework amounts to finding clusters of individuals satisfying these three desirable properties that we will denote throughout the paper:

(A) Heterotypy between species

(B) Homotypy within species

(M) Monophyly

We quickly observe that by definition, the phenotypic partition $\mathscr{P}$ satisfies the two desirable properties: (A) and (B). Unfortunately, phenotypically similar individuals may not be more genealogically related to one another than to any different individual. In other words $\mathscr{P}$ does not in general satisfy (M). Convergence and reversal events, by making a trait either appear several times independently in different parts of the tree or by making a trait disappear in subtrees, are classically invoked to explain the non-monophyly of $\mathscr{P}$. However, let us stress that even with traits evolving without convergence or reversal, individuals characterized by an ancestral phenotype may define a non-monophyletic subset of $\mathcal{X}$, a phenomenon called 'ancestral type retention' or 'plesiomorphy' (see Figure 1).

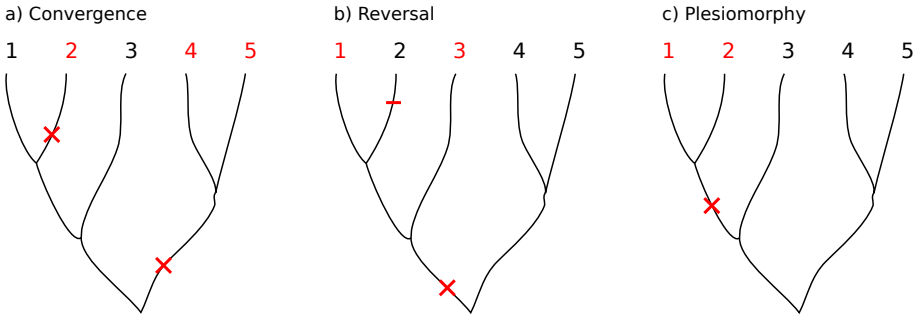

Fig. 1: Different scenarios leading to non-monophyletic phenotypic partitions, for the 'red' phenotype. a) The same phenotype arises twice independently in two branches. b) A new phenotype arises, and disappears later. c) The group of individuals showing the ancestral phenotype is not monophyletic.

The paper is organized as follows. In Section 1, we review the main species definitions used in the context of individual-based modeling of macro-evolution. In Section 2, we introduce the formalism required to study species partitions, and we make some preliminary observations on the three desirable properties. In Section 3, we prove that it mathematically makes sense to define the finest species partition satisfying (A) and (M) and the coarsest species partition satisfying (B) and (M). We call these respectively the *loose* and the *lacy* species partitions. Finally, we discuss the relevance of these definitions from both empirical and theoretical points of view.

1 Five species definitions in individual-based models

We provide here an overview of five modes of speciation that have been proposed so far in the context of individual-based models of diversification (see Kopp 2010 for a review and Figure 2 for illustration). Among these five modes, only the second one is intended to model specifically the geographical isolation of two subpopulations. The four other modes focus on modeling sympatric speciation by means of gradual accumulation of mutations.

Speciation by point mutation. This mode of speciation was proposed in the original framework of the NTB (Hubbell 2001; Jabot and Chave 2009). Differentiation occurs as the product of neutral mutations modeled by a point process on the genealogy. Each mutation confers a new type to the lineage carrying it (infinite-allele model) and to its descent before any new mutation

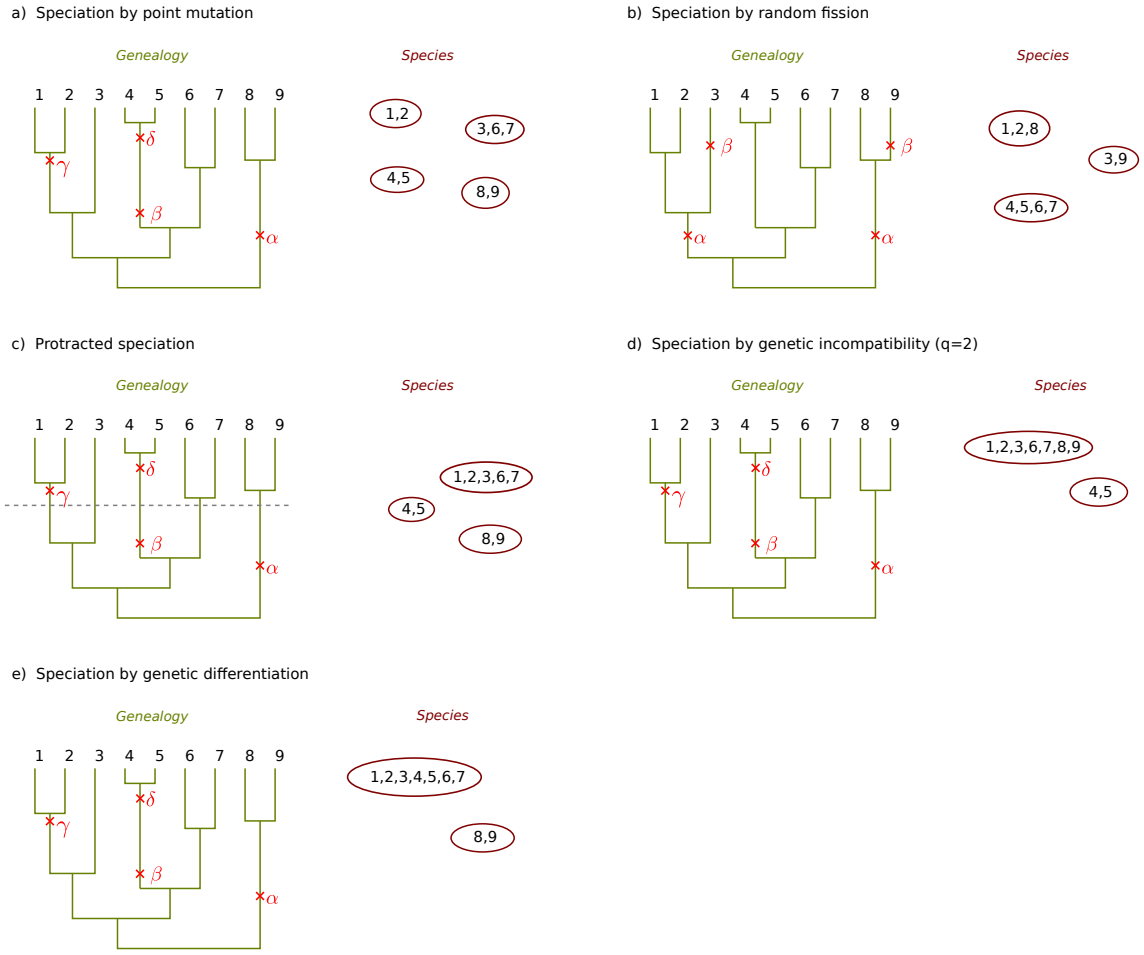

Fig. 2: The five modes of speciation proposed in individual-based models of macro-evolution. In each panel, the genealogy (tree) of individuals (integer labels) is given on the left, along with mutations (crosses) that confer new types (Greek letters) to individuals. The corresponding species partition is represented on the right of each panel (subsets of labels circled).

arises downstream. Species are then defined as groups of individuals carrying the same type. The phenotypic partition and the species partition thus coincide by definition.

Speciation by random fission, or peripheral isolates. These two closely related models have also been proposed first in the framework of the NTB (Hubbell 2001, 2003), but see also Lambert and Ma (2015). In these models, independently of the genealogy, each phenotypic class of individuals, interpreted as a geographic deme, may split at random times into two new demes. In Figure 2b, this is illustrated as mutations hitting simultaneously several lineages in the same phenotypic class, which endows them with the same new phenotype. The two models differ only

with regard to the size of the newly formed deme, whether the split is even (random fission) or uneven (peripheral isolate).

Protracted speciation. This model intends to reflect the general idea that speciation is not instantaneous (Rosindell et al. 2010; Lambert et al. 2015; Etienne et al. 2014). The differentiation process is usually assumed to be differentiation by point mutation under the infinite-allele model. A new phenotypic class is called an incipient species but becomes a so-called good species only after a fixed or random time duration. In other words, two individuals belong to different species if they carry different phenotypes and if they diverged far enough in the past. In Figure 2c, the species arisen from the mutation labelled γ and δ are still incipient at present time. More complex models of protracted speciation feature several stages that incipient species have to go through before becoming good species.

Speciation by genetic incompatibility. This generalization of the point mutation mode of speciation (Melián et al. 2012) is inspired by the model of Bateson-Dobzhansky-Muller incompatibilities (Orr 1995). Again, a first step consists in endowing the genealogy of individuals with neutral mutations. Then two individuals are said compatible if there are fewer than q mutations on the genealogical path linking them. Finally, species are the connected components of the graph associated to the compatibility relationship between individuals. For $q \neq 1$, there can be incompatible pairs of individuals in the same species, as can be seen in Figure 2d with individuals labelled 1 and 9 for example. Speciation by point mutation corresponds to the particular case $q = 1$.

Speciation by genetic differentiation. This model of speciation also assumes that phenotypic differentiation is driven by point mutations on the genealogy. Species are then defined as the smallest monophyletic groups of individuals such that any pair of individuals carrying the same phenotype are always in the same species (Manceau et al. 2015). We will show later that this definition, hereafter called *loose species definition*, always makes sense once given a phenotypic partition and a genealogy.

As can be seen in Figure 2, the first four models out of the five described in the previous section yield partitions of individuals into species that are in general non-monophyletic with respect to the

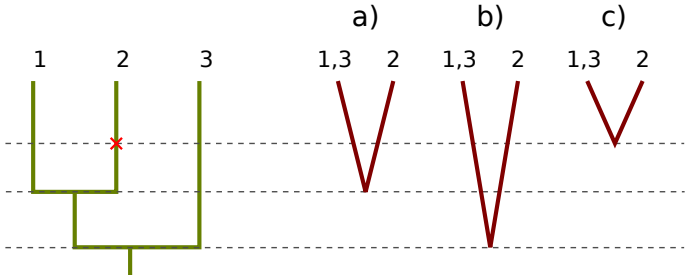

Fig. 3: Building the phylogeny out of the genealogy. Left panel: A fixed genealogy with one mutational event giving rise to a derived character responsible for partitioning $\{1, 2, 3\}$ into the two distinct phenotypic groups $\{1, 3\}$ and $\{2\}$, assumed to be different species. Right panel: Phylogenies associated with three possible choices of divergence times, from left to right: a) Shortest or b) Longest coalescence time between individuals of different species; c) Date of origin of the derived character.

underlying genealogy. This is problematic when it comes to measuring the phylogenetic relationship between species, reflecting their shared evolutionary history (Velasco 2008). In particular based on the true genealogy, there are multiple, arbitrary ways of defining the divergence time between two non-monophyletic species. We illustrate three of them in Figure 3.

The first two possibilities consist in relying on a time of divergence between individuals of the newly derived species and individuals of the ancestral, mother, species. One could imagine taking the shortest (scenario a), as well as the longest (scenario b), time of coalescence between these two groups of individuals. Among these two, scenario a) is the preferred option among population geneticists, for it provides a phylogeny consistent with the genealogy of alleles. A second possibility, and the one which is in ordinary usage in NTB-based studies (Jabot and Chave 2009), is to consider the date of appearance of the derived character as the time of divergence between the two species. We now argue that none of these possibilities is really appropriate from an evolutionary point of view.

Let us show for more clarity what these three possibilities lead to, in a scenario with three species (See Fig. 4). While scenarios a) and b) provide an unnecessary multifurcating tree, scenario c) even leads to a phylogeny topology distinct from what we would expect from the topology of the genealogy. Any representation of a phylogeny for non-monophyletic species would suffer such inconsistencies with the genealogy. To the contrary, the phylogeny of monophyletic species can be

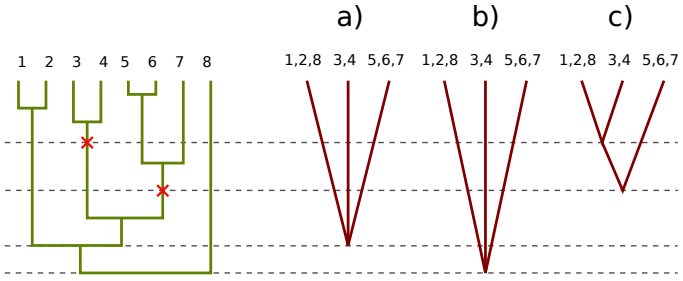

Fig. 4: Building the phylogeny out of the genealogy with three species. Left panel: A fixed genealogy with mutational events giving rise to three species. Right panel: Phylogenies associated with three possible choices of divergence times, from left to right: a) Shortest or b) Longest coalescence time between individuals of the ancestral and the derived species; c) Date of origin of the derived character.

obtained from the genealogy by considering as phylogenetic nodes all genealogical nodes that are most recent common ancestors between present-day species. As part of the current effort to bridging the gap between micro- and macro-evolution (Graham and Fine 2008; Rosindell et al. 2011; Pennell and Harmon 2013), it appears crucial to interlock the genealogical and phylogenetic scales. This is the rationale behind considering monophyly as a desirable property of species constructions.

In Sections 2 and 3, we will consider as given the genealogical tree T of the set of all present-day organisms $\mathcal{X}$, and the partition $\mathscr{P}$ of $\mathcal{X}$ into phenotypic groups. The tree T may have been generated under any model of population dynamics and the partition $\mathscr{P}$ may have been produced by any process of differentiation unfolding through time. With these data at hand, we formalize the three desirable properties of the species partition mentioned in the introduction and then study different ways to fulfill them.

2 Three desirable properties of species definitions

For each internal node of T, by a slight abuse of terminology, we call *clade* the subset of $\mathcal{X}$ comprising exactly all tips descending from this node. We denote by $\mathscr{H}$ *the collection of all clades of* T. Note that as a subset of $\mathcal{X}$, $\mathcal{X}$ itself is an element of $\mathscr{H}$, and that for every $x \in \mathcal{X}$, the singleton $\{x\}$ is an element of $\mathscr{H}$. Moreover, any two clades C and D elements of $\mathscr{H}$, are always either nested or mutually exclusive, meaning that $C \cap D$ can only be equal to C, D or $\varnothing$. Mathematically, a collection of nonempty subsets of $\mathcal{X}$ satisfying these properties is called a *hierarchy*, and it can be

shown that to any hierarchy corresponds a unique rooted tree with tips labelled by $\mathscr{X}$. Therefore, we will equivalently speak of T or of its hierarchy $\mathscr{H}$. For a nice discussion around the notion of hierarchy and neighboring concepts, see Steel (2014).

One should keep in mind that $\mathscr{H}$ and $\mathscr{P}$ are both collections of subsets of $\mathscr{X}$, but that $\mathscr{H}$ is not a partition. With this formalism, we define the species problem as : Given $\mathscr{H}$ and $\mathscr{P}$ find a partition $\mathscr{S}$ of $\mathscr{X}$, called the *species partition*, whose elements are called *species clusters* or simply *species*, satisfying one or more of the following three desirable properties:

(A) *Heterotypy between species.* Individuals in different species are phenotypically different, i.e. for each phenotypic cluster $P \in \mathscr{P}$ and for each species cluster $S \in \mathscr{S}$, either $P \subseteq S$ or $P \cap S = \varnothing$;

(B) *Homotypy within species.* Individuals in the same species are phenotypically identical, i.e. for each phenotypic cluster $P \in \mathscr{P}$ and for each species cluster $S \in \mathscr{S}$, either $S \subseteq P$ or $P \cap S = \varnothing$.

(M) *Monophyly.* Each species is a clade of T, i.e. $\mathscr{S} \subseteq \mathscr{H}$;

As mentioned in the introduction, if $\mathscr{S}$ satisfies both (A) and (B), then it is immediate from the preceding definitions that $\mathscr{S} = \mathscr{P}$. If in addition $\mathscr{S}$ satisfies (M) then $\mathscr{P} = \mathscr{S} \subseteq \mathscr{H}$. We record this as a first observation.

Observation 1 *Unless we are given $\mathscr{P}$ and $\mathscr{H}$ such that $\mathscr{P} \subseteq \mathscr{H}$ (that is, each phenotypic cluster is a clade in the first place), no species partition satisfies simultaneously (A), (B) and (M).*

Then our next question is: 'Is there a species partition $\mathscr{S}$ for which two of them hold?' For X, Y equal to A, B or M, we will write (XY) the property (X AND Y).

We already saw that $\mathscr{S} = \mathscr{P}$ satisfies (AB). Now let us go for species partitions $\mathscr{S}$ satisfying (M). To fulfill (A), each $S \in \mathscr{S}$ must contain all the phenotypic clusters it intersects. So in particular the partition $\mathscr{S}_1 := \{\mathscr{X}\}$ fulfills (AM). This trivial solution corresponds to assigning all the individuals of $\mathscr{X}$ to one single species. Symmetrically, to fulfill (B), each $S \in \mathscr{S}$ must be contained in all the phenotypic groups it intersects. So in particular the partition $\mathscr{S}_0$ made of all singletons fulfills (BM). This trivial solution corresponds to assigning each individual of $\mathscr{X}$ to a different species. This is recorded in the following observation.

Observation 2 *For any $\mathscr{P}$ and $\mathscr{H}$ and for any two desirable properties among (A), (B) and (M), there is at least one species partition $\mathscr{S}$ satisfying both properties.*

The species partitions $\mathscr{S}_1$ and $\mathscr{S}_0$ are obviously not biologically relevant. In particular, we would like to find species partitions that are *finer* than assigning all individuals to one single species, and *coarser* than assigning each individual to a different species.

We use the standard notions of finer and coarser partitions of a set (Bóna 2011). Let $\mathscr{S}$ and $\mathscr{S}'$ be two partitions of the set $\mathcal{X}$. We say that $\mathscr{S}$ is finer than $\mathscr{S}'$, and we write $\mathscr{S} \leq \mathscr{S}'$ if for each $S \in \mathscr{S}$ and each $S' \in \mathscr{S}'$, either $S \subseteq S'$ or $S \cap S' = \varnothing$. If $\mathscr{S} \leq \mathscr{S}'$, we say equivalently that $\mathscr{S}$ is finer than $\mathscr{S}'$ or that $\mathscr{S}'$ is coarser than $\mathscr{S}$.

Remark that two species partitions $\mathscr{S}$ and $\mathscr{S}'$ cannot always be compared, in the sense that they can satisfy neither $\mathscr{S} \leq \mathscr{S}'$ nor $\mathscr{S}' \leq \mathscr{S}$. The relation $\leq$ is thus not a linear order on all the partitions of $\mathcal{X}$, but is known to be a partial order (see Appendix A for details).

Now observe that properties (A) and (B) can precisely be stated in terms of inequalities associated with the partial order $\leq$ as follows.

Observation 3 *Consider a given phenotypic partition $\mathscr{P}$ and a species partition $\mathscr{S}$.*
$\mathscr{S}$ satisfies (A) if and only if $\mathscr{P} \leq \mathscr{S}$, and $\mathscr{S}$ satisfies (B) if and only if $\mathscr{S} \leq \mathscr{P}$.
As a consequence, if $\mathscr{S}_A$ satisfies (A) and $\mathscr{S}_B$ satisfies (B), then

$$\mathscr{S}_B \ \leq \ \mathscr{P} \ \leq \ \mathscr{S}_A$$

This leads us to investigate the possibility of defining 'the finest partition satisfying (AM)' as well as 'the coarsest partition satisfying (BM)'.

3 The lacy and loose species definitions

In general, there is no guarantee that the coarsest or finest partition of a given collection of partitions does belong to this collection. In particular, there is no guarantee that the coarsest (resp. finest) partition satisfying (BM) (resp. (AM)), does itself satisfy (BM) (resp. (AM)). However, we state the following result that ensures the existence of the finest partition satisfying (AM) and the coarsest partition satisfying (BM).

Theorem 1 *Given $\mathscr{P}$ and $\mathscr{H}$, there exists a unique finest partition of $\mathcal{X}$ satisfying (AM), and a unique coarsest partition of $\mathcal{X}$ satisfying (BM).*

This result is proved in Appendix B. It allows us to highlight and name two new different species definitions:

The *loose species definition* is the finest partition satisfying (AM).

The *lacy species definition* is the coarsest partition satisfying (BM).

For any species partition $\mathscr{S}$ satisfying (M), there is a unique phylogenetic tree $T_{\mathscr{S}}$ which represents the evolutionary relationships between the species in $\mathscr{S}$ consistently with the genealogy T. For every species S, since S is monophyletic there is a unique internal node $u(S)$ of T such that S is exactly constituted of the labels of the tips subtended by $u(S)$. Then $T_{\mathscr{S}}$ is obtained from T by merging, for every species S, the subtree descending from $u(S)$ into a single edge. This is expressed in terms of hierarchies in the following observation.

Observation 4 *Consider a given genealogical hierarchy $\mathscr{H}$ and species partition $\mathscr{S}$ satisfying (M). The hierarchy $\mathscr{H}_{\mathscr{S}}$ corresponding to the phylogenetic tree $T_{\mathscr{S}}$ can be defined by*

$$\mathscr{H}_{\mathscr{S}} := \{H \in \mathscr{H} : \ \exists S \in \mathscr{S}, \ S \subseteq H\}.$$

So for both the loose and the lacy species partition, there is a phylogeny consistent with the genealogy. Figure 5 shows both the lacy phylogeny and the loose phylogeny associated with a simple genealogy and a simple phenotypic partition.

For any genealogy T and phenotypic partition $\mathscr{P}$, we now describe a procedure to get the phylogeny corresponding either to the lacy or to the loose definition, without requiring the knowledge of species partitions. Interestingly, building $\mathscr{H}_{\mathscr{S}}$ this way offers a quick way to get $\mathscr{S}$ under the lacy and loose definitions, because species are the smallest sets of labels in $\mathscr{H}_{\mathscr{S}}$. The different steps of the algorithm are explained hereafter, illustrated in Figure 6 and formalized in Appendix C.

First, we classify all interior nodes of the genealogy as *convergent node* or *divergent node*. An interior node is *convergent* if there are at least two tips, one in each of its two descending subtrees, carrying the same phenotype. Otherwise the node is said to be *divergent*. Note that convergent nodes may be ancestors of divergent nodes when the phenotypic partition is not monophyletic. Second, we build a phylogeny by deciding which interior nodes are *phylogenetic nodes*, that is, appear in the phylogeny.

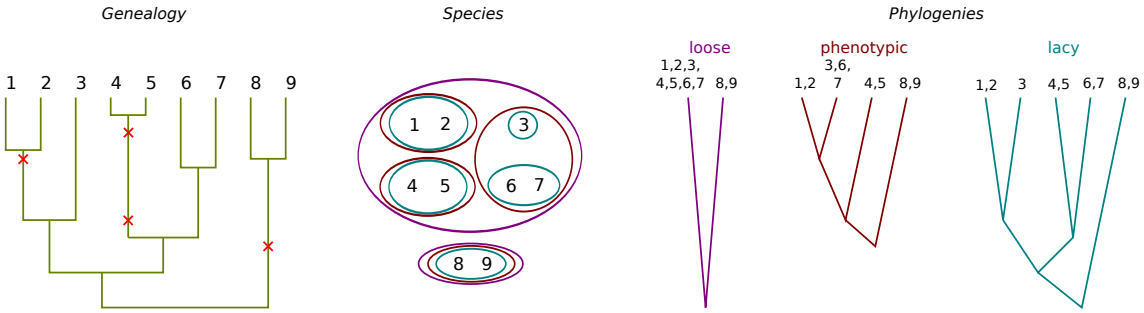

Fig. 5: Species partitions associated to each of three definitions. Left panel: The fixed genealogy with point mutations (infinite-allele model) leading to the phenotypic partition $\mathscr{P} = \{\{1,2\},\{3,6,7\},\{4,5\},\{8,9\}\}$. Middle panel: Inclusion relations between the three species partitions, as discussed in Observation 3, the loose partition is coarser than the phenotypic partition, which is coarser than the lacy partition. Right panel: Phylogenies corresponding to the three species partitions, from left to right: loose, phenotypic (under the arbitrary convention that divergence times are taken as mutation times), lacy.

Observation 5 *The loose phylogeny is obtained by declaring non-phylogenetic (i) all convergent nodes and (ii) all divergent nodes descending from a convergent node. Other nodes are declared phylogenetic.*

The lacy phylogeny is obtained by declaring phylogenetic (i) all divergent nodes and (ii) all convergent nodes ancestral to divergent nodes. Other nodes are declared non-phylogenetic.

The last observation holds due to the following reasons:

By definition, the two clades C and C' subtended by a convergent node satisfy $C \cap P \neq \varnothing$ and $C' \cap P \neq \varnothing$ for some phenotypic cluster $P \in \mathscr{P}$. As a consequence, these two clades have to be included in the same species cluster in a species partition satisfying the heterotypy property (A), that is, a convergent node cannot appear in a phylogeny satisfying (A). Conversely, any phylogeny whose nodes are included in the set of divergent nodes of the genealogy satisfies (A). The finest partition satisfying (A) corresponds to the phylogeny containing the largest number of divergent nodes, and only divergent nodes, as in the construction of the loose phylogeny proposed in the observation.

Symmetrically, for the two clades C, C' subtended by a divergent node, we have that $C \cap P \neq \varnothing$ implies $C' \cap P = \varnothing$ for any phenotypic cluster $P \in \mathscr{P}$. As a consequence, these two clades have to belong to two different species clusters in a species partition satisfying the homotypy property (B),

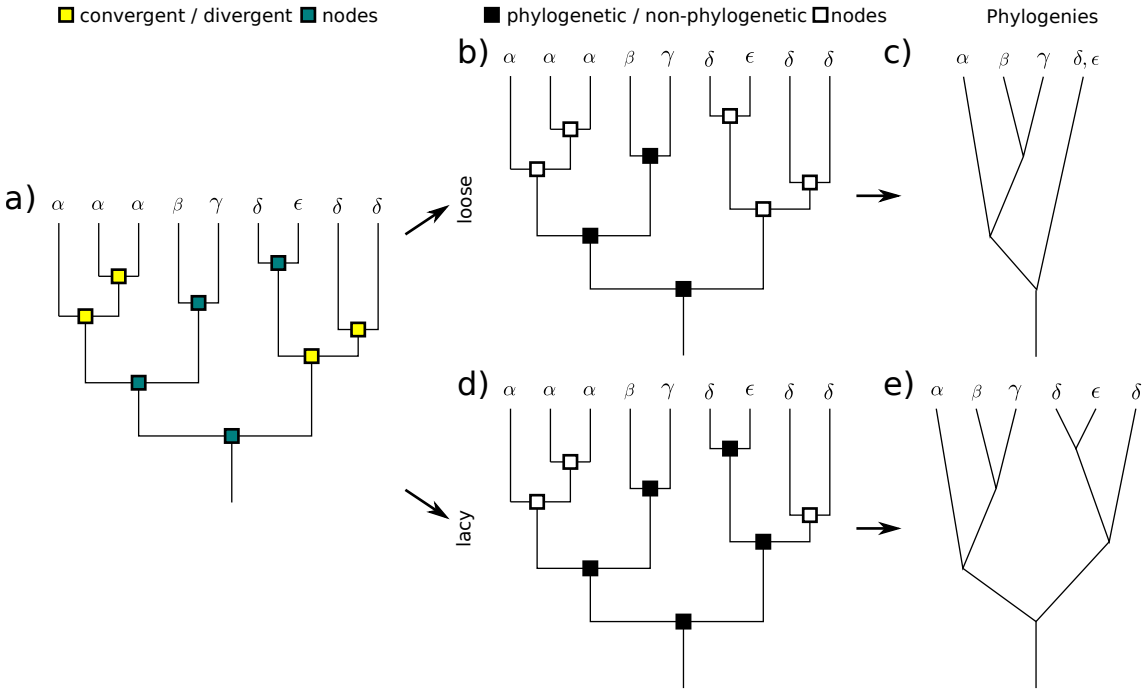

Fig. 6: Construction of the phylogeny under the lacy and loose species definitions. Greek letters correspond to phenotypes. a) The genealogy with interior nodes classified as convergent (light yellow) or divergent (dark blue). bd) The genealogy with interior nodes classified as non-phylogenetic (white) or phylogenetic (black). ce) The corresponding phylogeny. b) Loose: Light nodes and dark nodes descending from light nodes are colored white. d) Lacy: Dark nodes and light nodes ancestors of a dark node are colored black.

that is, any divergent node has to appear in a phylogeny satisfying (B). Conversely, any phylogeny whose nodes contain all divergent nodes of the genealogy satisfies (B). The coarsest partition satisfying (B) corresponds to the phylogeny containing the smallest number of convergent nodes, but all divergent nodes, as in the construction of the lacy phylogeny proposed in the observation.

Discussion

The present study builds on recent attempts to describe evolutionary trees on various scales simultaneously. The multi-species coalescent is one of the most influential of these attempts (Maddison 1997). In this model, a species tree is first specified and given the species tree, the gene genealogies are drawn from a censored coalescent (i.e., ancestral lineages can coalesce only if they lie in the same ancestral species). In particular, this approach has been used to assess the relevance of the reciprocal monophyly criterion to recognize species (Hudson and Coyne 2002; Mehta et al. 2016).

Note that this framework introduces a top-down coupling between the macro-evolutionary scale and the micro-evolutionary scale, thus relying on an external species definition. In contrast, the bottom-up approach that we adopted consists in assuming that macro-evolutionary patterns are shaped by micro-evolutionary processes. Let us point out two studies similar in spirit to ours, which make several proposals to lump together lower-order taxa (e.g. species) in order to build trees on higher-order taxa (e.g. genera) (Aldous et al. 2008, 2011). The authors ground their definitions on the knowledge of a phylogeny with point mutations, the main differences being that each branching event distinguishes a mother and a daughter lineage, and that monophyly is not a desirable property in their work.

On the contrary, we put to the forefront the monophyly property (M), together with (A) heterotypy between species and (B) homotypy within species. We explicitly defined and compared three species definitions, each of these satisfying a different set of properties: Phenotypic (AB), Loose (AM) and Lacy (BM).

Additionally, we stress that the most popular way of defining species in individual-based modeling studies of macro-evolution (i.e., speciation by point mutation) in general leads to non-monophyletic species. In contrast, the loose species definition, previously used in the context of macro-evolution (Manceau et al. 2015), systematically yields monophyletic species. Here we extended this study and compared it to a third species definition also satisfying (M), the lacy species definition. Finally, we provided a standardized procedure to build the lacy and loose species partitions given a genealogy and a phenotypic partition.

In practice, the task of systematists is the inference of ancestral relationships between individual organisms from molecular sequence and phenotype data and the characterization of species from those data. Classifying diversity is notoriously difficult for many reasons, including the difficulty of choosing the appropriate level of description, the ubiquitous presence of convergent evolution and reversal events, and the difficulty to agree on a unique species concept (Mayden 1997; De Queiroz 2007; Baum 2009).

On the other hand, the task of modelers, assuming a fully known individual-based evolutionary history, appears at first sight trivial. They face, however, the same difficulty in defining a proper

species concept. Even within the very simple framework that we considered, three distinct species definitions came out. They all fit the general species definition of *'separately evolving metapopulation lineages'* (De Queiroz 2007), while satisfying distinct desirable properties. We argue that comparing species definitions based on the properties they fulfill in simple models might help shed light on the species problem.

Let us draw parallels between our theoretical considerations and the habits of taxonomists. Practically, the sole phenotypic information is usually sufficient do decide whether a taxon above the species level is monophyletic. And indeed in systematics, monophyly has long been a criterion for defining taxa above the species level, even before the rise of molecular methods. On the contrary, phenotypic information alone is certainly not sufficient to diagnose monophyletic species. The use of molecular markers has brought the question of intra-species monophyly (M) to the forefront. Today, it is standard to use multiple sequence alignments to automatically delineate putative species: from a single-locus phylogeny (Fujisawa and Barraclough 2013), from multiple gene trees (Yang and Rannala 2010), or from the raw alignment (Puillandre et al. 2012). Species descriptions based on these methods are thus more likely to concern monophyletic groups of individuals than earlier. This recent requirement for species monophyly puts taxonomists in front of new dilemmas. Loose or lacy? Crudely speaking, one could say that 'splitter' taxonomists more often lean for the lacy definition, while 'lumper' taxonomists are more willing to use the loose definition. More precisely, the diagnosis of species as phenotypically homogeneous groups of individuals that can be separated solely on a molecular basis into what is known as cryptic species (Bickford et al. 2007) corresponds to the lacy definition. On the contrary, taxonomists preferring to ensure that species are diagnosable and monophyletic units, two properties stressed as 'priority taxon naming criteria' (Vences et al. 2013) use the loose definition.

Note that advanced theoretical work has been undertaken in a context of sexually-reproducing organisms (Dress et al. 2010; Kwok 2011; Alexander 2013; Alexander et al. 2015). While we based our study on the knowledge of only one genealogy, even the genealogical history of supposedly 'asexual' real-world organisms such as bacteria shows evidence for horizontal gene transfer events (Puigbò et al. 2013). The genealogical history of organisms should in general be represented as a

non-tree network, or as a collection of gene genealogies, making far more complex the question of grouping individuals into taxa (Hudson and Coyne 2002; Samadi and Barberousse 2006). Their framework is closer to biological reality, but much less connected to most modeling studies in macro-evolution.

Individual-based modeling is a promising avenue for understanding macro-evolution from first principles, as it may allow evolutionary biologists to describe explicitly the stochastic demography of whole metacommunities and the ecological interactions between different types of individuals in each community. We believe that these processes may have left enough signal in both the shape of evolutionary trees and the patterns of contemporary biodiversity, so as to be unraveled by statistical inference. Understanding how species, the elementary units of macro-evolution, are formed and deformed by these processes remains a major challenge, to which the present work hopefully contributes.

References

Aguilée, R., A. Lambert, and D. Claessen. 2011. Ecological speciation in dynamic landscapes. J. Evolution. Biol. 24:2663–2677.

Aldous, D., M. Krikun, and L. Popovic. 2008. Stochastic models for phylogenetic trees on higher-order taxa. J. Math. Biol. 56:525–557.

Aldous, D. J., M. A. Krikun, and L. Popovic. 2011. Five statistical questions about the tree of life. Syst. Biol. 60:318–328.

Alexander, S. A. 2013. Infinite graphs in systematic biology, with an application to the species problem. Acta Biotheor. 61:181–201.

Alexander, S. A., A. de Bruin, and D. J. Kornet. 2015. An alternative construction of internodons: The emergence of a multi-level tree of life. B. Math. Biol. 77:23–45.

Avise, J. C. and R. M. Ball. 1990. Principles of genealogical concordance in species concepts and biological taxonomy. Oxford Surv. Evol. Bio. 7:45–67.

Baum, D. A. 2009. Species as ranked taxa. Syst. Biol. 58:74–86.

Bickford, D., D. J. Lohman, N. S. Sodhi, P. K. Ng, R. Meier, K. Winker, K. K. Ingram, and I. Das. 2007. Cryptic species as a window on diversity and conservation. Trends Ecol. Evol. 22:148–155.

Bock, W. J. 2004. Species: the concept, category and taxon. J. Zool. Syst. Evol. Res. 42:178–190.

Bóna, M. 2011. A walk through combinatorics: an introduction to enumeration and graph theory. World scientific.

De Queiroz, K. 2007. Species concepts and species delimitation. Syst. Biol. 56:879–886.

De Queiroz, K. and M. J. Donoghue. 1988. Phylogenetic systematics and the species problem. Cladistics 4:317–338.

Dress, A., V. Moulton, M. Steel, and T. Wu. 2010. Species, clusters and the 'tree of life': A graph-theoretic perspective. J. Theor. Biol. 265:535–542.

Durrett, R. 2008. Probability models for DNA sequence evolution. Springer.

Etienne, R. S., H. Morlon, and A. Lambert. 2014. Estimating the duration of speciation from phylogenies. Evolution 68:2430–2440.

Fujisawa, T. and T. G. Barraclough. 2013. Delimiting species using single-locus data and the generalized mixed yule coalescent (GMYC) approach: a revised method and evaluation on simulated datasets. Syst. Biol. 62:707–724.

Gascuel, F., R. Ferrière, R. Aguilée, and A. Lambert. 2015. How ecology and landscape dynamics shape phylogenetic trees. Syst. Biol. 64:590–607.

Graham, C. H. and P. V. A. Fine. 2008. Phylogenetic beta diversity: Linking ecological and evolutionary processes across space in time. Ecol. Lett. 11:1265–1277.

Hennig, W. 1965. Phylogenetic systematics. Annu. Rev. Entomol. 10:97–116.

Hubbell, S. P. 2001. The Unified Neutral Theory of Biodiversity and Biogeography. Princeton University Press, Princeton.

Hubbell, S. P. 2003. Modes of speciation and the lifespans of species under neutrality: a response to the comment of Robert E. Ricklefs. Oikos 100:193–199.

Hudson, R. R. and J. A. Coyne. 2002. Mathematical consequences of the genealogical species concept. Evolution 56:1557–1565.

Jabot, F. and J. Chave. 2009. Inferring the parameters of the neutral theory of biodiversity using phylogenetic information and implications for tropical forests. Ecol. Lett. 12:239–248.

Kopp, M. 2010. Speciation and the neutral theory of biodiversity. Bioessays 32:564–570.

Kwok, R. B. H. 2011. Phylogeny, genealogy and the linnaean hierarchy: a logical analysis. J. Math. Biol. 63:73–108.

Lambert, A. and C. Ma. 2015. The coalescent in peripatric metapopulations. J. Appl. Probab. 52:538–557.

Lambert, A., H. Morlon, and R. S. Etienne. 2015. The reconstructed tree in the lineage-based model of protracted speciation. J. Math. Biol. 70:367–397.

Maddison, W. P. 1997. Gene trees in species trees. Syst. Biol. 46:523–536.

Manceau, M., A. Lambert, and H. Morlon. 2015. Phylogenies support out-of-equilibrium models of biodiversity. Ecol. Lett. 18:347–356.

Mayden, R. L. 1997. A hierarchy of species concepts: the denouement in the saga of the species problem. *in* Species, the units of biodiversity. Claridge, M.F. and Dawah, H.A. and Wilson, M. R.

Mehta, R. S., D. Bryant, and N. A. Rosenberg. 2016. The probability of monophyly of a sample of gene lineages on a species tree. Proc. Natl. Acad. Sci. U. S. A. 113:8002–8009.

Melián, C. J., D. Alonso, S. Allesina, R. S. Condit, and R. S. Etienne. 2012. Does sex speed up evolutionary rate and increase biodiversity? PLoS Comput. Biol. 8:1–9.

Missa, O., C. Dytham, and H. Morlon. 2016. Understanding how biodiversity unfolds through time under neutral theory. Phil. Trans. R. Soc. B 371.

Morlon, H. 2014. Phylogenetic approaches for studying diversification. Ecol. Lett. 17:508–525.

Orr, H. A. 1995. The population genetics of speciation: the evolution of hybrid incompatibilities. Genetics 139:1805–1813.

Pennell, M. W. and L. J. Harmon. 2013. An integrative view of phylogenetic comparative methods: Connections to population genetics, community ecology, and paleobiology. Ann. NY Acad. Sci. 1289:90–105.

Puigbò, P., Y. I. Wolf, and E. V. Koonin. 2013. Seeing the tree of life behind the phylogenetic forest. BMC Biol. 11:1–3.

Puillandre, N., A. Lambert, S. Brouillet, and G. Achaz. 2012. ABGD, automatic barcode gap discovery for primary species delimitation. Mol. Ecol. 21:1864–1877.

Pyron, R. A. and F. T. Burbrink. 2013. Phylogenetic estimates of speciation and extinction rates for testing ecological and evolutionary hypotheses. Trends Ecol. Evol. 28:729–736.

Regan, C. T. 1925. Organic evolution. Nature 116:398–401.

Rosindell, J., S. J. Cornell, S. P. Hubbell, and R. S. Etienne. 2010. Protracted speciation revitalizes the neutral theory of biodiversity. Ecol. Lett. 13:716–727.

Rosindell, J., L. J. Harmon, and R. S. Etienne. 2015. Unifying ecology and macroevolution with individual-based theory. Ecol. Lett. 18:472–482.

Rosindell, J., S. P. Hubbell, and R. S. Etienne. 2011. The unified neutral theory of biodiversity and biogeography at age ten. Trends Ecol. Evol. 26:340–348.

Samadi, S. and A. Barberousse. 2006. The tree, the network, and the species. Biol. J. Linn. Soc. 89:509–521.

Sneath, P. H. A. 1976. Phenetic taxonomy at the species level and above. Taxon 25:437–450.

Stadler, T. 2013. Recovering speciation and extinction dynamics based on phylogenies. J. Evolution. Biol. 26:1203–1219.

Steel, M. 2014. Tracing evolutionary links between species. Am. Math. Mon. 121:771–792.

Velasco, J. D. 2008. Species concepts should not conflict with evolutionary history, but often do. Stud. Hist. Philos. Biol. Biomed. Sci. 39:407–414.

Vences, M., J. M. Guayasamin, A. Miralles, and I. De La Riva. 2013. To name or not to name: Criteria to promote economy of change in linnaean classification schemes. Zootaxa 3636:201–244.

Yang, Z. and B. Rannala. 2010. Bayesian species delimitation using multilocus sequence data. Proc. Natl. Acad. Sci. U. S. A. 107:9264–9269.

Appendix
The species problem from the modeler's point of view

Some of the results stated in Sections A and B are classical results in combinatorics for partially ordered sets (see Bóna 2011, chapter 16). For the sake of self-containment and because all readers may not be familiar with these notions, we nevertheless expose them here.

A 'Finer than', a partial order relation on $\mathcal{X}$-partitions

Definition 1 Let $\mathscr{S}_1$ and $\mathscr{S}_2$ be two $\mathcal{X}$-partitions. We say that $\mathscr{S}_1$ is finer than $\mathscr{S}_2$, and we write $\mathscr{S}_1 \leq \mathscr{S}_2$ if $\forall S_1 \in \mathscr{S}_1, \forall S_2 \in \mathscr{S}_2,\ S_1 \cap S_2 \in \{\varnothing, S_1\}$.

We detail here the three criteria that make the 'finer than' relation a partial order on the set of $\mathcal{X}$-partitions.

Proof One must check the reflexivity, antisymmetry and transitivity properties.

- Reflexivity. Take any $\mathcal{X}$-partition $\mathscr{S}$. Then for all $S_1, S_2 \in \mathscr{S}$ we either have $S_1 \cap S_2 = S_1$ if $S_1 = S_2$, or $S_1 \cap S_2 = \varnothing$ otherwise. It follows that $\mathscr{S} \leq \mathscr{S}$.
- Antisymmetry. Take two $\mathcal{X}$-partitions denoted $\mathscr{S}_1$ and $\mathscr{S}_2$, verifying $\mathscr{S}_1 \leq \mathscr{S}_2$ and $\mathscr{S}_2 \leq \mathscr{S}_1$. Then for all $(S_1, S_2) \in \mathscr{S}_1 \times \mathscr{S}_2,\ S_1 \cap S_2 \in \{\varnothing, S_1\}$ and $S_1 \cap S_2 \in \{\varnothing, S_2\}$.
  If $S_1 \cap S_2 \neq \varnothing$, it follows that $S_1 = S_2$, and finally $\mathscr{S}_1 = \mathscr{S}_2$.
- Transitivity. Take now three $\mathcal{X}$-partitions denoted $\mathscr{S}_1, \mathscr{S}_2, \mathscr{S}_3$, verifying $\mathscr{S}_1 \leq \mathscr{S}_2$ and $\mathscr{S}_2 \leq \mathscr{S}_3$. Let $S_1 \in \mathscr{S}_1$ and $S_3 \in \mathscr{S}_3$ and assume that $S_1 \cap S_3 \neq \varnothing$. Then there is $x \in S_1 \cap S_3$ and we let S_2 be the unique element of $\mathscr{S}_2$ such that $x \in S_2$. Thus $S_1 \cap S_2 \neq \varnothing$ and $S_2 \cap S_3 \neq \varnothing$, which implies by assumption that $S_2 \cap S_1 = S_1$ and $S_2 \cap S_3 = S_2$. So we see that $S_1 \subseteq S_2 \subseteq S_3$, so that $S_1 \cap S_3 = S_1$. □

B Proof of Theorem 1

Here we will consider sets of partitions verifying one or two desirable properties. Hence the following definitions

$$\Sigma_A := \{\mathcal{X}\text{-partitions satisfying (A)}\}$$

$$\Sigma_B := \{\mathcal{X}\text{-partitions satisfying (B)}\}$$

$$\Sigma_M := \{\mathcal{X}\text{-partitions satisfying (M)}\}$$

$$\Sigma_{AM} := \{\mathcal{X}\text{-partitions satisfying (AM)}\} = \Sigma_A \cap \Sigma_M$$

$$\Sigma_{BM} := \{\mathcal{X}\text{-partitions satisfying (BM)}\} = \Sigma_B \cap \Sigma_M$$

$$\Sigma_{AB} := \{\mathcal{X}\text{-partitions satisfying (AB)}\} = \Sigma_A \cap \Sigma_B$$

We will see that the collection of $\mathcal{X}$-partitions Σ_M plays a singular role in Theorem 1. This is due to the characterization of Σ_M by the fact that there is a hierarchy $\mathscr{H}$ (here the hierarchy associated with the genealogy T) such that

$$\mathscr{S} \in \Sigma_M \iff \mathscr{S} \subseteq \mathscr{H}.$$

Also recall that the collections of $\mathcal{X}$-partitions Σ_A and Σ_B can be defined as follows

$$\mathscr{S} \in \Sigma_A \iff \forall P \in \mathscr{P},\ \forall S \in \mathscr{S},\ P \cap S \in \{\varnothing, P\}$$

$$\mathscr{S} \in \Sigma_B \iff \forall P \in \mathscr{P},\ \forall S \in \mathscr{S},\ P \cap S \in \{\varnothing, S\}.$$

In this section, we aim at giving a proof of Theorem 1 which can now be restated as follows

$$\exists \mathscr{S}_{\text{loose}} \in \Sigma_{AM},\ \text{such that } \forall \mathscr{S} \in \Sigma_{AM},\ \mathscr{S}_{\text{loose}} \leq \mathscr{S}$$

$$\exists \mathscr{S}_{\text{lacy}} \in \Sigma_{BM},\ \text{such that } \forall \mathscr{S} \in \Sigma_{BM},\ \mathscr{S} \leq \mathscr{S}_{\text{lacy}}$$

The proof is divided into two parts. First, given a set of partitions Σ (resp. $\Sigma \subseteq \Sigma_M$), we prove the existence of the finest (resp. coarsest) partition finer (resp. coarser) than any element of Σ, which we call $\inf \Sigma$ (resp. $\sup \Sigma$). Second, we show that $\inf \Sigma_{AM} \in \Sigma_{AM}$ and $\sup \Sigma_{BM} \in \Sigma_{BM}$, hence yielding the definitions $\mathscr{S}_{\text{loose}} := \inf \Sigma_{AM}$ and $\mathscr{S}_{\text{lacy}} := \sup \Sigma_{BM}$.

B.1 Defining the supremum and the infimum of a set of $\mathcal{X}$-partitions

Definition 2 For any non-empty collection Σ of $\mathcal{X}$-partitions, we define the two relations $\underline{\mathcal{R}}_\Sigma$ and $\overline{\mathcal{R}}_\Sigma$ on $\mathcal{X}$ by

$$\forall (x, y) \in \mathcal{X}^2,\ x\ \underline{\mathcal{R}}_\Sigma\ y \iff \forall \mathscr{S} \in \Sigma,\ \exists S \in \mathscr{S},\ x \in S \text{ and } y \in S$$

$$x\ \overline{\mathcal{R}}_\Sigma\ y \iff \exists \mathscr{S} \in \Sigma,\ \exists S \in \mathscr{S},\ x \in S \text{ and } y \in S.$$

Lemma 1 *For any non-empty collection Σ of $\mathcal{X}$-partitions, $\underline{\mathcal{R}}_\Sigma$ is an equivalence relation. For any non-empty collection Σ of $\mathcal{X}$-partitions such that $\Sigma \subseteq \Sigma_M$, $\overline{\mathcal{R}}_\Sigma$ is an equivalence relation.*

Proof The *reflexivity* and *symmetry* of the two relations are easily seen. Now let us prove their transitivity. Let Σ be a non-empty collection of $\mathcal{X}$-partitions, and $(x, y, z) \in \mathcal{X}^3$ such that $x\ \underline{\mathcal{R}}_\Sigma\ y$ and $y\ \underline{\mathcal{R}}_\Sigma\ z$. Let $\mathscr{S} \in \Sigma$. By definition,

$$\exists S_1 \in \mathscr{S},\ x \in S_1 \text{ and } y \in S_1$$

$$\exists S_2 \in \mathscr{S},\ y \in S_2 \text{ and } z \in S_2$$

It follows that $y \in S_1 \cap S_2$, and because $\mathscr{S}$ is a partition, $S_1 = S_2$. Finally, with $S := S_1 = S_2$, there exists $S \in \mathscr{S}$ such that $x \in S$ and $z \in S$, so that $x\ \underline{\mathcal{R}}_\Sigma\ z$ and we can conclude that $\underline{\mathcal{R}}_\Sigma$ is transitive.

Now let $\Sigma \subseteq \Sigma_M$ be a non-empty collection of $\mathcal{X}$-partitions and $(x, y, z) \in \mathcal{X}^3$ such that $x \overline{\mathcal{R}}_\Sigma y$ and $y \overline{\mathcal{R}}_\Sigma z$. By definition,

$$\exists \mathscr{S}_1 \in \Sigma,\ \exists S_1 \in \mathscr{S}_1,\ x \in S_1 \text{ and } y \in S_1$$

$$\exists \mathscr{S}_2 \in \Sigma,\ \exists S_2 \in \mathscr{S}_2,\ y \in S_2 \text{ and } z \in S_2$$

Because $\Sigma \subseteq \Sigma_M$, $\mathscr{S}_1 \subseteq \mathscr{H}$ and $\mathscr{S}_2 \subseteq \mathscr{H}$, so that $S_1 \in \mathscr{H}$ and $S_2 \in \mathscr{H}$. From the definition of hierarchy, we get $S_1 \cap S_2 \in \{\varnothing, S_1, S_2\}$. Since $y \in S_1 \cap S_2$, we have $S_1 \cap S_2 \neq \varnothing$.
Suppose that $S_1 \cap S_2 = S_2$. It follows that $\exists \mathscr{S}_1 \in \Sigma,\ \exists S_1 \in \mathscr{S}_1,\ x \in S_1$ and $z \in S_1$.
Suppose that $S_1 \cap S_2 = S_1$. It follows that $\exists \mathscr{S}_2 \in \Sigma,\ \exists S_2 \in \mathscr{S}_2,\ x \in S_2$ and $z \in S_2$. So $x \overline{\mathcal{R}}_\Sigma z$ and we can conclude that $\overline{\mathcal{R}}_\Sigma$ is transitive. □

Definition 3 For any non-empty collection Σ of $\mathcal{X}$-partitions, we call $\inf \Sigma$ the $\mathcal{X}$-partition induced by the equivalence relation $\underline{\mathcal{R}}_\Sigma$. For any non-empty collection Σ of $\mathcal{X}$-partitions such that $\Sigma \subseteq \Sigma_M$, we call $\sup \Sigma$ the $\mathcal{X}$-partition induced by the equivalence relation $\overline{\mathcal{R}}_\Sigma$.

Readers familiar with lattice theory will note that these definitions match the usual 'meet' and 'join' operators used for lattices, and in particular the lattice of partitions of a set, ordered by refinement. For the other readers, recall first that any equivalence relation on a set $\mathcal{X}$ induces an $\mathcal{X}$-partition obtained by placing all elements in relation in one cluster. Further, the following lemma justifies the notation inf and sup.

Lemma 2 *Let Σ be any non-empty collection of $\mathcal{X}$-partitions. Then for any $\mathscr{S} \in \Sigma$, $\inf \Sigma \leq \mathscr{S}$. Let Σ be any non-empty collection of $\mathcal{X}$-partitions such that $\Sigma \subseteq \Sigma_M$. Then for any $\mathscr{S} \in \Sigma$, $\mathscr{S} \leq \sup \Sigma$.*

Proof Let Σ be any non-empty collection of $\mathcal{X}$-partitions and $S \in \inf \Sigma$. Let also $\mathscr{S} \in \Sigma$ and $S' \in \mathscr{S}$. We need to prove that $S \cap S' \in \{\varnothing, S\}$. Assume that $S \cap S' \neq \varnothing$ and $S \cap S' \neq S$. Then there is $x \in S \cap S'$ and $y \in S$ such that $y \notin S'$. Because $x, y \in S$, by definition of $\inf \Sigma$, we have $x \underline{\mathcal{R}}_\Sigma y$ and by definition of $\underline{\mathcal{R}}_\Sigma$, $\exists S'' \in \mathscr{S},\ x, y \in S''$. So S' and S'' are both elements of $\mathscr{S}$ containing x, which implies that $S' = S''$ and contradicts $y \notin S'$.

Now let Σ be any non-empty collection of $\mathcal{X}$-partitions such that $\Sigma \subseteq \Sigma_M$ and $S \in \sup \Sigma$. Let also $\mathscr{S} \in \Sigma$ and $S' \in \mathscr{S}$. We need to prove that $S \cap S' \in \{\varnothing, S'\}$. Assume that $S \cap S' \neq \varnothing$ and $S \cap S' \neq S'$. Then there is $x \in S \cap S'$ and $y \in S'$ such that $y \notin S$. Because $x \in S$ and $y \notin S$, by definition of $\sup \Sigma$, x and y are not in relation $\overline{\mathcal{R}}_\Sigma$ and by definition of $\overline{\mathcal{R}}_\Sigma$, either $x \notin S'$ or $y \notin S'$ and we get a contradiction. □

Note that, in general, we can have $\inf \Sigma \notin \Sigma$ and $\sup \Sigma \notin \Sigma$. Here are two examples to provide the reader with some intuition.

*Example 1.*Take

$$\mathcal{X} = \{1, 2, 3, 4\}$$

$$\mathscr{S} = \{\{1\}, \{2\}, \{3, 4\}\}$$

$$\mathscr{S}' = \{\{1, 2\}, \{3\}, \{4\}\}$$

$$\Sigma = \{\mathscr{S}, \mathscr{S}'\}$$

In this case, we get $\inf \Sigma = \{\{1\}, \{2\}, \{3\}, \{4\}\}$, which does not belong to Σ. Moreover, if we define the hierarchy $\mathscr{H} := \{\{1, 2, 3, 4\}, \{1, 2\}, \{3, 4\}, \{1\}, \{2\}, \{3\}, \{4\}\}$, we have $\Sigma \subseteq \Sigma_M$, which allows us to consider $\sup \Sigma = \{\{1, 2\}, \{3, 4\}\}$ which again does not belong to Σ.

*Example 2.*Take

$$\mathcal{X} = \{1, 2, 3, 4\}$$

$$\mathscr{S} = \{\{1, 3, 4\}, \{2\}\}$$

$$\mathscr{S}' = \{\{1, 2\}, \{3, 4\}\}$$

$$\Sigma = \{\mathscr{S}, \mathscr{S}'\}$$

In this case, we get $\inf \Sigma = \{\{1\}, \{2\}, \{3, 4\}\}$, which does not belong to Σ. Moreover, there is no $\mathcal{X}$-hierarchy $\mathscr{H}$ such that $\mathscr{S}, \mathscr{S}' \in \mathscr{H}$. Then we can see that the relation $\overline{\mathcal{R}}_\Sigma$ is not an equivalence relation on $\mathcal{X}$, because $1 \; \overline{\mathcal{R}}_\Sigma \; 2$ and $1 \; \overline{\mathcal{R}}_\Sigma \; 3$, but we do not have $2 \; \overline{\mathcal{R}}_\Sigma \; 3$. Thus, $\sup \Sigma$ is not defined.

B.2 Proving that $\inf \Sigma_{AM} \in \Sigma_{AM}$ and $\sup \Sigma_{BM} \in \Sigma_{BM}$

In order to prove that $\inf \Sigma_{AM} \in \Sigma_{AM}$ and $\sup \Sigma_{BM} \in \Sigma_{BM}$, we will rely on properties of $\inf \Sigma$ and $\sup \Sigma$ presented in the following lemma.

Lemma 3 *For any non-empty collection Σ of $\mathcal{X}$-partitions, for any $S \in \inf \Sigma$, S can be written in the form of the following non-empty intersection*

$$S = \bigcap_{\mathscr{S} \in \Sigma : S \subseteq S^* \in \mathscr{S}} S^* \tag{S1}$$

For any non-empty collection Σ of $\mathcal{X}$-partitions such that $\Sigma \subseteq \Sigma_M$, for any $S \in \sup \Sigma$, S can be written in the form of the following non-empty union

$$S = \bigcup_{\mathscr{S} \in \Sigma : S \supseteq S^* \in \mathscr{S}} S^* \tag{S2}$$

In addition,

$$\exists \mathscr{S} \in \Sigma, \; \exists S^* \in \mathscr{S}, \; S^* = S. \tag{S3}$$

Proof We begin with proving (S1). Let Σ be any non-empty collection of $\mathcal{X}$-partitions and consider $S \in \inf \Sigma$. Now set

$$S' := \bigcap_{\mathscr{S} \in \Sigma : S \subseteq S^* \in \mathscr{S}} S^*$$

and let us prove that $S = S'$. According to Lemma 2 that for any $\mathscr{S} \in \Sigma$, $\inf \Sigma \leq \mathscr{S}$ so $\exists! S^* \in \mathscr{S}$ such that $S \subseteq S^*$. This proves that the intersection in the definition of S' is not empty. Now by definition of S' we have $S \subseteq S'$, which also implies $S' \neq \varnothing$. We need to show now that $S' \subseteq S$. Let x be any element of S' and y be any element of S. Then for any $\mathscr{S} \in \Sigma$, there is (a unique) $S^* \in \mathscr{S}$ such that $S \subseteq S^*$ and by definition of S', we have $x \in S^*$. But since $S \subseteq S^*$ we also have $y \in S^*$. This shows that for any $\mathscr{S} \in \Sigma$ there is $S^* \in \mathscr{S}$ such that $x \in S^*$ and $y \in S^*$. This can be expressed equivalently as $x \underline{\mathcal{R}}_\Sigma y$, so that x and y are in the same element of $\inf \Sigma$, that is $x \in S$.

Now let us prove (S2). Let Σ be any non-empty collection of $\mathcal{X}$-partitions such that $\Sigma \subseteq \Sigma_M$ and let $S \in \sup \Sigma$. Set

$$S' := \bigcup_{\mathscr{S} \in \Sigma : S \supseteq S^* \in \mathscr{S}} S^*$$

and let us prove that $S = S'$. According to Lemma 2 that for all $\mathscr{S} \in \Sigma$, $\mathscr{S} \leq \sup \Sigma$, so $\exists S^* \in \mathscr{S}$ such that $S^* \subseteq S$. In particular, the intersection in the definition of S' is not empty and $S' \neq \varnothing$. Now by definition of S' we have $S' \subseteq S$. We need to show now that $S \subseteq S'$. Let x be any element of S and y be any element of S'. Since $S' \subseteq S$, $y \in S$ so that x and y are in the same element of $\sup \Sigma$, which can be expressed equivalently as $x \overline{\mathcal{R}}_\Sigma y$. Now by definition of $\overline{\mathcal{R}}_\Sigma$, there is $\mathscr{S} \in \Sigma$ and $S^* \in \mathscr{S}$ such that $x, y \in S^*$. Now since $S^* \cap S \neq \varnothing$, we have $S^* \subseteq S$, which shows by definition of S' that $x \in S'$.

It remains to show (S3)

$$\exists \mathscr{S} \in \Sigma,\ \exists S^* \in \mathscr{S},\ S^* = S.$$

Let us prove by induction on $n \geq 1$ that for any $F \subseteq S$ of cardinality n, there is $\mathscr{S} \in \Sigma$ and $S^* \in \mathscr{S}$ such that $F \subseteq S^* \subseteq S$. The result will follow by taking $F = S$. For $n = 1$, the property holds due to to (S2). Let $n \geq 1$ strictly smaller than the cardinality of S and assume that the property holds for all integers smaller than or equal to n. Let F be any subset of S of cardinality $n+1$ and write $F = F_1 \cup \{x\}$, where $x \notin F_1$. Since F_1 is of cardinality n there is $\mathscr{S}_1 \in \Sigma$ and $S_1 \in \mathscr{S}_1$ such that $F_1 \subseteq S_1 \subseteq S$. Let $y \in F_1$. There is also $\mathscr{S}_2 \in \Sigma$ and $S_2 \in \mathscr{S}_2$ such that $\{x, y\} \subseteq S_2 \subseteq S$. Now because $\mathscr{S}_1, \mathscr{S}_2 \in \Sigma \subseteq \Sigma_M$, we have $\mathscr{S}_1, \mathscr{S}_2 \subseteq \mathscr{H}$, so that $S_1 \in \mathscr{H}$ and $S_2 \in \mathscr{H}$. From the definition of hierarchy, we get $S_1 \cap S_2 \in \{\varnothing, S_1, S_2\}$. Since $y \in S_1 \cap S_2$, we have $S_1 \cap S_2 \neq \varnothing$, so one of the two, denoted S^* contains the other one. In particular, $F_1 \subseteq S^* \subseteq S$ and $\{x, y\} \subseteq S^* \subseteq S$, which shows that $F = F_1 \cup \{x\} \subseteq S^* \subseteq S$ and terminates the proof. □

We can now go back to the proof of Theorem 1.

Proof (i) $\inf \Sigma_{AM} \in \Sigma_A$: Consider $S \in \inf \Sigma_{AM}$ and $P \in \mathscr{P}$. From Lemma 3, we get

$$S \cap P = P \cap \left(\bigcap_{\mathscr{S} \in \Sigma_{AM} : S \subseteq S^* \in \mathscr{S}} S^* \right) = \bigcap_{\mathscr{S} \in \Sigma_{AM} : S \subseteq S^* \in \mathscr{S}} (P \cap S^*)$$

Now for each $S^* \in \mathscr{S} \in \Sigma_{AM}$, $P \cap S^* \in \{\varnothing, P\}$, thus leading to $S \cap P \in \{\varnothing, P\}$, that is $\inf \Sigma_{AM} \in \Sigma_A$.

(ii) $\inf \Sigma_{AM} \in \Sigma_M$: Consider $S \in \inf \Sigma_{AM}$. From Lemma 3, we get

$$S = \bigcap_{\mathscr{S} \in \Sigma_{AM} : S \subseteq S^* \in \mathscr{S}} S^*$$

Now for each $S^* \in \mathscr{S} \in \Sigma_{AM}$, $S^* \in \mathscr{H}$. Moreover, the hierarchy $\mathscr{H}$ is closed under finite, non-disjoint intersections, thus leading to $S \in \mathscr{H}$, that is $\inf \Sigma_{AM} \in \Sigma_M$.

(iii) $\sup \Sigma_{BM} \in \Sigma_B$: Consider $S \in \sup \Sigma_{BM}$ and recall from Lemma 3 that there is $\mathscr{S} \in \Sigma_{BM}$ and $S^* \in \mathscr{S}$ such that $S = S^*$. Now for any $P \in \mathscr{P}$,

$$S \cap P = S^* \cap P \in \{\varnothing, S^*\} = \{\varnothing, S\},$$

so that $\sup \Sigma_{BM} \in \Sigma_B$.

(iv) $\sup \Sigma_{BM} \in \Sigma_M$: Consider $S \in \sup \Sigma_{BM}$ and $S^* = S$ as previously. Since $S^* \in \mathscr{H}$, $S \in \mathscr{H}$, so that $\sup \Sigma_{BM} \in \Sigma_M$.

This shows that $\inf \Sigma_{AM} \in \Sigma_{AM}$ and $\sup \Sigma_{BM} \in \Sigma_{BM}$, which completes the proof of Theorem 1. □

C Construction of the lacy and loose phylogenies

This section aims at formalizing mathematically the construction of the lacy and loose phylogenies presented in the main text.

Recall that an interior node is convergent if there are two tips, one in each of its two descending subtrees, carrying the same phenotype, otherwise this node is said to be divergent. We will say that the two clades subtended by a convergent (resp. divergent) node are convergent (resp. divergent). We define $\mathscr{H}_d$ as the collection of divergent clades, that is

$$\mathscr{H}_d = \{h \in \mathscr{H} : \exists h', h'' \in \mathscr{H},\ h' = h \cup h'',\ \forall P \in \mathscr{P},\ h \cap P = \varnothing \text{ or } h'' \cap P = \varnothing\} \cup \mathcal{X}$$

We similarly consider phylogenetic and non-phylogenetic clades for either the loose or the lacy definition. We call $\mathscr{H}_{\text{loose}}$ and $\mathscr{H}_{\text{lacy}}$ the collection of phylogenetic clades for the loose and lacy definitions respectively. The procedure described in the main text amounts to defining

$$\mathscr{H}_{\text{loose}} = \mathscr{H} \backslash \{h \in \mathscr{H} : \exists h_c \in \mathscr{H} \backslash \mathscr{H}_d,\ h \subseteq h_c\}$$

$$\mathscr{H}_{\text{lacy}} = \{h \in \mathscr{H} : \exists h' \in \mathscr{H},\ h \cup h' \in \mathscr{H},\ h \cap h' \in \{\emptyset, h'\},\ \exists h_d \in \mathscr{H}_d,\ h_d \subseteq h'\}$$