Version dated: August 24, 2016

RH: THE SPECIES PROBLEM IN MODELING STUDIES

The species problem from the modeler's point of view

Marc Manceau[1,3,4], Amaury Lambert[2,3]

[1]*Muséum National d'Histoire Naturelle, 75005 Paris, France;*

[2]*Laboratoire Probabilités et Modèles Aléatoires, UPMC Univ Paris 06, 75005 Paris, France;*

[3]*Center for Interdisciplinary Research in Biology, Collège de France, CNRS UMR 7241, 75005 Paris, France;*

[4]*Institut de Biologie, École Normale Supérieure, CNRS UMR 8197, 75005 Paris, France*

Corresponding author: Marc Manceau, Institut de Biologie de l'École Normale Supérieure, 46 rue d'Ulm, 75005 Paris, France; E-mail: marc.manceau@ens.fr

Abstract.— How to define and delineate species is a long-standing question sometimes called the species problem. In modern systematics, species should be groups of individuals sharing characteristics inherited from a common ancestor which distinguish them from other such groups. A good species definition should thus satisfy the following three desirable properties: (A) Heterotypy between species, (B) Homotypy within species and (E) Exclusivity, or monophyly, of each species.

In practice, systematists seek to discover the very traits for which these properties are satisfied, without the a priori knowledge of the traits which have been responsible for differentiation and speciation nor of the true ancestral relationships between individuals. Here to the contrary, we focus on individual-based models of macro-evolution, where both the differentiation process and the population genealogies are explicitly modeled, and we ask: How and when is it possible, with this significant information, to delineate species in a way satisfying most or all of the three desirable properties (A), (B) and (E)?

Surprisingly, despite the popularity of this modeling approach in the last two decades, there has been little progress or agreement on answers to this question. We prove that the three desirable properties are not in general satisfied simultaneously, but that any two of them can. We show mathematically the existence of two natural species partitions: the finest partition satisfying (A) and (E) and the coarsest partition satisfying (B) and (E). For each of them, we propose a simple algorithm to build the associated phylogeny. We stress that these two procedures can readily be used at a higher level, namely to cluster species into monophyletic genera.

The ways we propose to phrase the species problem and to solve it should further refine models and our understanding of macro-evolution.
(Keywords: Gene Genealogy, Phylogeny, Microevolution, Individual-Based, Species Concept, Macroevolution, Infinite-Allele Model, Neutral Biodiversity Theory)

The problem of agreeing on what should be considered as the biologically most relevant concept of species and of constructing a general procedure to assign dead or alive organisms to the right species is a long-standing question called the *species problem*. The species problem is both a conceptual question (to define the species concept) and a practical problem (how to classify individuals into species) which is not restricted to species but concerns all taxonomic levels - even if the need for precision or robustness may be less pronounced for intermediate levels (e.g., families) (Bock 2004).

Classification problems can be found in all areas of science and often prove to be hard statistical ones. In systematics in particular, individuals can be clustered according to some selection of characteristics (morphology, behavior, genotype, specific associations of these...) and the partitions obtained for several different characteristics have great chances to be incompatible (De Queiroz 2007). Several of the most notable evolutionary biologists (e.g. Darwin, Dobzhansky, Mayr, Simpson, Hennig...) took a stand on the species problem leading to often vigorous debates

(see Mayden 1997; Mallet 2001; De Queiroz 2005 for overviews of historical disagreements).

In particular, the foundation and spread of cladistics by Hennig (1965) marked a radical change of paradigm in the systematic classification (De Queiroz and Donoghue 1988). Species are meant to be groups of individuals sharing derived characteristics, i.e. characteristics inherited from a common ancestor which distinguish them from other such groups. In other words, defining species, or any higher-level taxon, amounts to finding characteristics (called synapomorphies), such that (i) any two individuals in distinct putative species can be distinguished based on one characteristic, (ii) individuals belonging to the same putative species share one characteristic and (iii) the characteristic is inherited from a common ancestor. In the rest of the paper we will denote these three properties as follows:

(i) Heterotypy between species: (A)

(ii) Homotypy within species: (B)

(iii) Exclusivity: (E)

and call them the *three desirable properties* (A), (B) and (E). We emphasize the subtle difference between the term 'monophyly', which refers to the entire descendance of a given ancestor, and the term 'exclusivity', which refers to the group of extant descendants of this ancestor (Velasco 2009).

In practice, the task of systematists is to infer the ancestral relationships between individual organisms from gene sequence and phenotype data and to characterize species from those phenotypes which are synapomorphies. Recently, several approaches have been designed to help delineate putative species from sequence data: from a single phylogeny (Fujisawa and Barraclough 2013; Zhang et al. 2013), from gene trees (Yang and Rannala 2010), or from raw molecular data (Puillandre et al. 2012). The leading idea of all these methods is to search for a gap separating small intra-species variability and large inter-species variability. In this note, we do not wish to discuss the existing methods used in practice to delineate and identify species. Rather, we concentrate on the following, related theoretical question:

> *How* and *when* is it possible to delineate species, in a way satisfying some or all of the three desirable properties, in an ideal situation where the characteristics are specified

and *the entire genealogy is known*?

Formally, this question is 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).

The first reason why we assume that characteristics are specified and the genealogy is known is to question the existence of theoretical limits, and if so, identify them, to our ability to solve the species problem, independently of the degree of information available. Second, in the last two decades, a popular way of studying macro-evolution has consisted in performing computer-intensive simulations of individual-based stochastic processes of species diversification (Jabot and Chave 2009; Pigot et al. 2010; Aguilée et al. 2011, 2013; Rosindell et al. 2015; Gascuel et al. 2015; Missa et al. 2016). In such individual-based simulations, the population dynamics and the associated genealogy are explicitly modeled, as well as the process of appearance of new characteristics potentially giving rise to new species. When building the model, it is crucial to decide how to cluster individuals into species in a way that is susceptible to satisfy most of the three desirable properties (A), (B) and (E). This crucial step gave its title to the paper ('the species problem from the modeler's point of view'). As we will see, there is no natural way of achieving this step and different authors often choose different criteria. Further, we show in this paper that the three desirable properties cannot in general be simultaneously satisfied. However, for any two properties among the three, there are always solutions to the species problem that satisfy both. Even more precisely, we prove mathematically that there is always one natural candidate which satisfies (A) and (E) and one natural candidate which satisfies (B) and (E). The first one is called the 'loose species definition' and is the finest species partition satisfying (A) and (E). The second one is called the 'lacy species definition' and is the coarsest species partition satisfying (B) and (E).

The paper is organized as follows. In the next section, we fix some notation and explain why in general, the species clustering satisfying (A) and (B), called the phenotypic partition, does not satisfy (E). In a second section, we review the main species definitions used by modelers in the context of individual-based simulations of macro-evolutionary processes. In the third

section, we introduce a classical formalism used for the study of species partitions, and we make some preliminary observations on the three desirable properties. In particular, species partitions satisfying (B) are always finer than or equal to species partitions satisfying (A). In the fourth section, we specify when it makes sense mathematically to define the finest or the coarsest partition and show that in the present context we can use these notions to define the lacy and the loose species partitions. Finally, we discuss the relevance of these propositions from both empirical and theoretical points of view.

Paraphyly of phenotype-based partitions

A convenient way of representing genealogical relationships within a sample $\mathcal{X}$ of individual organisms is in the form of a tree, where each tip represents (and so 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. Trees are accurate descriptions of ancestral relationships in strictly asexual populations, for which each individual has only one parent.

To the contrary, sexually reproducing organisms show a much more complex genealogical history, that can still be represented as a network of past and present individuals (vertices) connected by parent-child relationships (edges). In this network, individuals may be connected by more than one path, and genetic information may split and follow several of these paths. The ancestral history of different parts of the genome may then be composed of many potentially incongruent trees. In the most spectacular cases, some individuals belonging to distant species can be more closely related at some loci than individuals belonging to sister species, a phenomenon called *incomplete lineage sorting* (Maddison 1997). Even the genealogical history of so-called 'asexual' organisms such as bacteria deviates from a strict tree due to horizontal gene transfer events (Puigbò et al. 2013).

Trees have also been used as simplified representations of these complex genealogical networks. Of particular interest, Dress et al. (2010) have formally described five different ways to make up genealogical trees from complex genealogical networks. For the sake of convenience, we will nonetheless make the assumption of asexual reproduction throughout this paper. In the

following, the genealogical history of the set $\mathcal{X}$ will thus be represented by a rooted tree denoted T.

Organisms are endowed with an innumerable quantity of measurable characteristics. Different individuals can show different genotypes, different morphologies, different behaviors, different geographic ranges and so forth. We will assume to have chosen a specific subset of characteristics, that we will subsume under the word *phenotype*. Then all individuals in $\mathcal{X}$ can be grouped into clusters of individuals showing the same phenotype. We will call the associated $\mathcal{X}$-partition the *phenotypic partition*, denoted $\mathscr{P}$.

By definition, the phenotypic partition $\mathscr{P}$ satisfies the two desirable properties (A) Heterotypy between species and (B) Homotypy within species. Unfortunately, phenotypically similar individuals may not be more related to one another than to any different individual, in other words $\mathscr{P}$ does not in general satisfy (E) Exclusivity. This situation, which is called *paraphyly* of $\mathscr{P}$ with respect to T, can be the result of different mechanisms :

Convergence. A phenotypic trait appears several times independently in different parts of the tree (Fig. 1, left panel),

Reversal. A phenotypic trait appears once and disappears later in one or several subtrees (Fig. 1, middle panel),

Ancestral type retention. Individuals with the ancestral phenotype may define a non-exclusive subset of $\mathcal{X}$ (Fig. 1, right panel).

Convergence and reversal events are collectively designated under the term of *homoplasy*. In view of the previous discussion and Figure 1 (left and middle panels), characteristics experiencing homoplasies must be excluded when defining species in the cladistic fashion. On the contrary, with homoplasy-free processes of phenotypic evolution, the groups of individuals carrying the derived characteristic are always exclusive. However, it is very important at this stage to note that because of ancestral type retention, even homoplasy-free evolution can still lead to the paraphyly of the phenotypic partition $\mathscr{P}$ as a whole, as can be seen on the right panel of Figure 1.

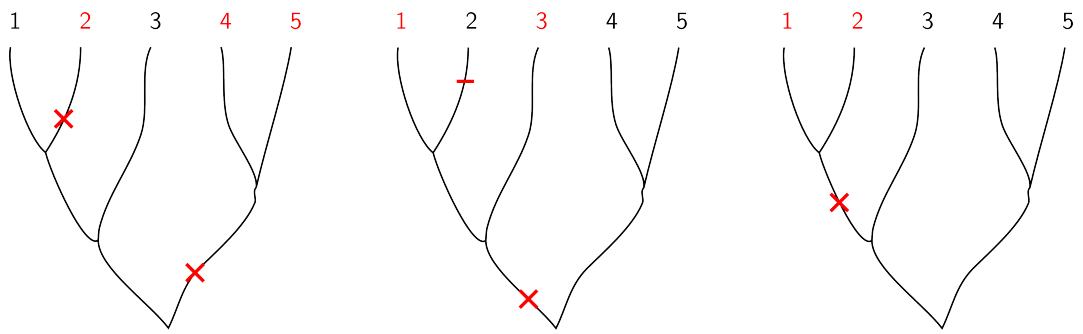

FIGURE 1: Different scenarios leading to paraphyletic phenotypic partitions, for the 'red' phenotype. Left panel, convergence: The same phenotype arises twice independently in two branches. Middle panel, reversal: A new phenotype arises, and disappears later. Right panel, ancestral type retention: The group of individuals showing the ancestral phenotype is not exclusive.

In the next section, we review the different ways used in the literature to define species in individual-based models of species diversification where phenotypic evolution is assumed to be homoplasy-free.

SPECIES DEFINITIONS IN INDIVIDUAL-BASED MODELS

Macro (species) and micro (individual) modeling approaches

Mathematical models in macro-ecology and 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 entities, thought of as particles that can give birth to other particles (i.e., speciate) during a given lifetime (i.e., before extinction) (for reviews, see Stadler 2013; Pyron and Burbrink 2013; Morlon 2014). In contrast, empirical evolutionary processes (differentiation, reproduction, selection) are usually described at the level of individuals.

These processes, encompassed under the name of micro-evolution, are specifically modeled in the fields of population genetics and adaptive dynamics. The former field seeks to study the evolution of genetic polymorphism and the ancestral structure of populations, potentially taking into account a wealth of biological details such as the presence of recombination, of epistasis, of

different selection regimes, of spatial structure, of genetic incompatibilities and so forth (see Crow and Kimura 1970; Hartl and Clark 1997 for an introduction or Durrett 2008 for an overview of recent mathematical developments). The latter field focuses on the evolution of quantitative characters under the assumption of rare mutations. This framework puts emphasis on the details of ecological processes, which may be modeled explicitly. Contrary to population genetics, the fitness is not given *a priori* but emerges naturally from the dynamics of rare mutants in a resident population at equilibrium (Metz et al. 1996). Researchers have attempted to scale-up both population genetics and adaptive dynamics to the meso-evolutionary scale, i.e., the scale at which speciation occurs. In population genetics, speciation can be modeled by the accumulation of so-called Bateson-Dobzhansky-Muller incompatibilities (Nei et al. 1983; Orr 1995; Hudson and Coyne 2002). Adaptive dynamics have contributed to characterizing ecological conditions under which different phenotypes can appear and co-occur in a system (Geritz et al. 1997). The so-called branching phenomenon has offered theoretical grounds to sympatric speciations occurring in nature (Dieckmann and Doebeli 1999; Doebeli and Dieckmann 2003). However up until today, there is no clear understanding of how these two frameworks can be used to scale-up to the macro-evolutionary scale of diversification.

A third modeling approach, the Neutral Theory of Biodiversity (Hubbell 2001) opened a new way of thinking species in models of macro-ecology and macro-evolution. In this approach, births, deaths and extinction, differentiation and speciation, are described at the level of individuals, relying on three major steps: (i) First, the genealogy of individuals, represented in the form of a genealogical tree T, is produced under a given scenario of population dynamics (fixed or stochastic), assuming that each birth event and each death event target all individuals with the same probability (assumption of selective neutrality); (ii) Second, a process of phenotypic differentiation acting on the lineages of T generates a partition $\mathscr{P}$ of all individuals sampled at any given time (or at possibly different times) into phenotypic groups; (iii) A species definition is postulated, which is used to cluster individuals into different species, in relation to the genealogies and phenotypes generated in the first two steps. 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 of Step (i) has been modeled in different ways in previous studies. The most widely used models are the Wright-Fisher and Moran model (Durrett 2008) from population genetics, which assume a constant metapopulation size through time, and lead to a probabilistic representation of genealogies in the form of the well-known (structured) Kingman coalescent (Kingman 1982). The neutrality assumption ensures that the next two steps can be conducted independently of the first one. In the next section, we explain how these two steps have been treated in the literature, first the process of differentiation (ii), which leads to a partition of the population into phenotypic groups (see Kopp 2010 for an alternative review), and then the species definition (iii), which leads to a partition of individuals into species.

The five modes of speciation

To our knowledge, five modes of speciation have been proposed so far by theoreticians studying the properties of individual-based models of diversification. Among these five modes, only the second one is intended to model specifically the geographical isolation of two subpopulations, either with a random split (allopatric speciation) or with an uneven one (peripatric speciation). The four other modes focus on modeling sympatric speciation by means of gradual accumulation of mutations. These five propositions are illustrated in Figure 2 showing striking differences between all these definitions.

Speciation by point mutation. This mode of speciation was proposed in the original framework of the Neutral Theory of Biodiversity (Hubbell 2001; Chave 2004; Latimer et al. 2005; Jabot and Chave 2009; Davies et al. 2011). Differentiation occurs as the product of neutral mutations modeled by a Poisson point process on the genealogy. Each mutation confers a new type to the lineage carrying it (infinite-allele model) and to its descendance before any new mutation 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, but due to ancestral type retention, species may not be exclusive, as seen in Figure 2a with the group of individuals labelled $3, 6, 7$.

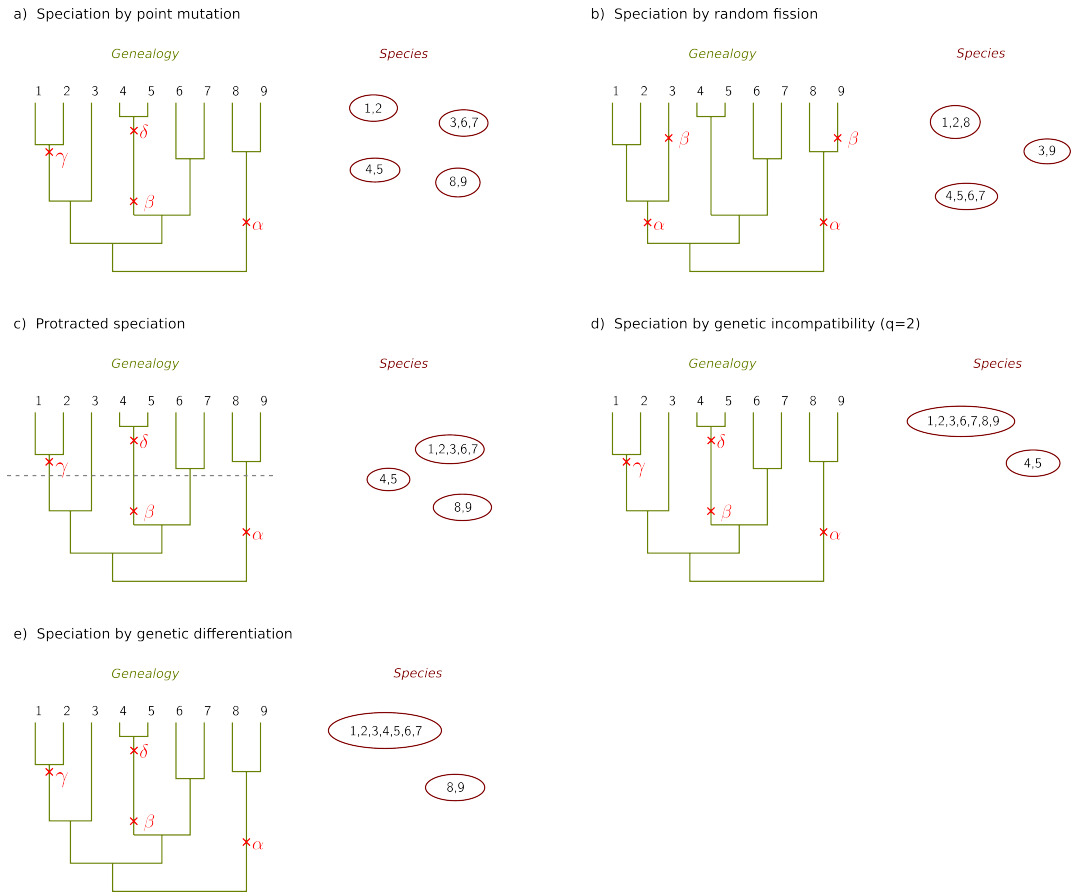

FIGURE 2: The five modes of speciation proposed in individual-based models of macro-evolution. In each panel, the genealogy (green tree) of individuals (integer labels) is given on the left, along with mutations (red crosses) that confer new types (greek letters) to individuals (infinite-allele model). The corresponding species partition is represented on the right of each panel (subsets of labels circled in red). a) *Speciation by point mutation.* Two individuals are in the same species if and only if they carry the same type. b) *Speciation by random fission.* The same mutation hits several individuals simultaneously and confers the same new type to their descents. Two individuals are in the same species if and only if they carry the same type. c) *Protracted speciation.* Two individuals are in the same species if their divergence/coalescence time is smaller than a given threshold (grey dashed line) or if they carry the same type. d) *Speciation by genetic incompatibility.* Two individuals are said compatible if there are less than q mutations they do not share; species are the connected components of the compatibility graph. e) *Speciation by genetic differentiation.* Species are the smallest exclusive groups of individuals, such that individuals carrying the same type always belong to the same group.

Speciation by random fission, or peripheral isolates. These two closely related models have also been proposed first in the framework of the Neutral Theory of Biodiversity (Hubbell 2003; Etienne and Haegeman 2011), 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 (newly formed deme). The two propositions differ only with regard to the distribution of the number of individuals belonging to the newly formed deme, whether it is smaller (i.e., a peripheral isolate) or whether the split is even (random fission). This model shares similarities with the multi-species coalescent model (Maddison 1997; Degnan and Rosenberg 2009) in the sense that the gene genealogy is embedded in a coarser tree, i.e., the species tree in the multi-species coalescent and the implicit history of successive fissions in the present model.

Protracted speciation. This model intends to reflect the general idea that speciation is not instantaneous (Rosindell et al. 2010; Etienne and Rosindell 2011; Lambert et al. 2015; Etienne et al. 2014). It is characterized by the definition it gives of species regardless of the differentiation process, which 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 have diverged far enough into the past. For example in Figure 2c, the species arisen from the mutation labelled c is 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 (De Aguiar et al. 2009; 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 the number of mutations they do not share is smaller than a fixed threshold q. 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. The point mutation mode of speciation corresponds to the particular case $q = 1$.

Speciation by genetic differentiation. This fifth model of speciation was proposed recently by Manceau et al. (2015). We assume given a phenotypic partition of individuals, generated for example by point mutations on the genealogy. Species are then defined as the smallest exclusive groups of individuals such that any pair of individuals in the same phenotypic group are always in the same species. We will show later that this definition, hereafter called 'loose species definition', always makes sense once given a phenotypic partition and a genealogy.

Building the phylogeny out of the 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 paraphyletic with respect to the underlying genealogy. In the context of macro-ecology, the existence of paraphyletic species does not come as a surprise. However, it becomes problematic when it comes to measuring the phylogenetic relationship between species, reflecting their shared evolutionary history. According to Velasco (2008), who called this issue 'the paraphyly problem', *'placing [non-exclusive groups] on the tips of trees misrepresents history and leads to incorrect inferences'*. In particular based on the true genealogy, there are multiple, arbitrary ways of defining the divergence time between two paraphyletic species, for example:

a) The shortest coalescence time between pairs of individuals belonging to the two species;

b) The longest coalescence time between pairs of individuals belonging to the two species;

c) The date of appearance of the derived character responsible for their differentiation.

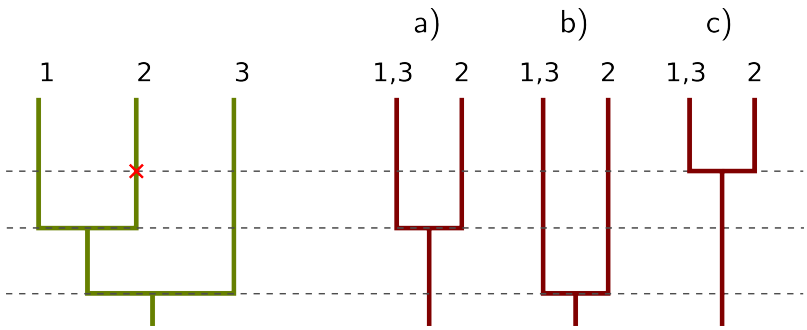

FIGURE 3: Building the phylogeny out of the genealogy. Left panel: A fixed genealogy with one mutational event giving rise to a derived character responsible for partitioning $\{1, 2, 3\}$ into the two distinct phenotypic groups $\{1, 3\}$ and $\{2\}$, assumed to be different species. Right panel: The 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.

Note that (c) was the choice made by Jabot and Chave (2009). As illustrated in Figure 3 on a toy example, none of these three solutions gives a reasonable account of the phylogenetic history of these species. Strangely enough, we have not found any previous discussion of this problem in the literature. As part of the current effort to bridging the gap between micro- and macro-evolution (Barraclough and Nee 2001; Estes and Arnold 2007; Graham and Fine 2008; Rosindell et al. 2011; Pennell and Harmon 2013), it becomes urgent to understand how the genealogical and the phylogenetic scales interact.

In the next section, we will consider as given data the evolutionary history of a metapopulation, represented in the form of a genealogical tree T, that has been generated under any model of population dynamics. We will also consider given a partition $\mathscr{P}$ into phenotypic groups of all individuals at present time, which has been produced by any process of differentiation unfolding through time on the genealogy, such as one of the five scenarios discussed above. In order to address the question of the species definition, we start by more precisely describing the three desirable properties of the species partition referred to earlier and then study different ways to fulfill them.

Three desirable properties of species definitions

In this section, we seek to formalize the three desirable properties of species definitions mentioned in the Introduction.

For each internal node of the genealogical tree 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 $\mathcal{X}$, the entire $\mathcal{X}$ 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 $\mathcal{X}$. 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).

From now on, we assume that we are given:

- The rooted genealogical tree T of the individuals labelled by the set $\mathcal{X}$ or, equivalently, the collection $\mathscr{H}$ of all clades of T;
- A phenotype-based partition of $\mathcal{X}$, denoted $\mathscr{P}$.

One should keep in mind that $\mathscr{H}$ and $\mathscr{P}$ are both collections of subsets of $\mathcal{X}$, but that $\mathscr{H}$ is not a partition. With this formalism, the species problem amounts to finding a partition $\mathscr{S}$ of $\mathcal{X}$, 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, meaning that 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, meaning that for each phenotypic cluster $P \in \mathscr{P}$ and for each species cluster $S \in \mathscr{S}$, either $S \subseteq P$ or $P \cap S = \varnothing$.

(E) *Exclusivity.* All species are exclusive, meaning that each species cluster is a clade of T, that is $\mathscr{S} \subseteq \mathscr{H}$;

Notice that if $\mathscr{S}$ satisfies both (A) and (B), then it is immediate from the preceding definitions that $\mathscr{S} = \mathscr{P}$. And if in addition $\mathscr{S}$ satisfies (E) then $\mathscr{P} = \mathscr{S} \subseteq \mathscr{H}$. We can record this in the following 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 (E).*

Recall from the first section of this paper that even under the very restricted assumption of a homoplasy-free phenotypic evolution, as is the case for point mutations in the infinite-allele model, ancestral type retention can cause paraphyletic phenotypic partitions (see Fig. 1 and 2). There is thus in general no species partition $\mathscr{S}$ for which the three desirable properties hold at the same time. 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 E, we will write (XY) for species satisfying both (X) and (Y).

Of course the phenotypic partition $\mathscr{S} = \mathscr{P}$ satisfies (AB). Now let us go for species partitions satisfying (E). Recall that a species partition $\mathscr{S}$ is exclusive if for any $S \in \mathscr{S}$, $S \in \mathscr{H}$. To fulfill (A), each $S \in \mathscr{S}$ must contain all the phenotypic clusters it intersects. So in particular the partition $\mathscr{S}_1 := \{\mathcal{X}\}$ fulfills (AE). This trivial solution corresponds to assigning all the individuals of the sample $\mathcal{X}$ to the same 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 (BE). This trivial solution corresponds to assigning each individual of the sample $\mathcal{X}$ to a different species. This can be recorded in the following trivial observation.

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

The species partitions $\mathscr{S}_1$ and $\mathscr{S}_0$ given previously as examples satisfying respectively (AE) and (BE) are in general not biologically relevant. In particular, we would like to find species partitions that are *finer* than assigning all individuals to one single species, or *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 $\mathscr{S}$.

Note 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 $\mathscr{S}' \leq \mathscr{S}$, as shows the next example on $\mathcal{X} = \{1, 2, 3, 4\}$:

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

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

In this example, $\mathscr{S}$ and $\mathscr{S}'$ cannot be compared. 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 SI for details).

Now recall the definitions of the properties (A) and (B) and observe that they 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 $\mathscr{S}_A$ is a species partition satisfying (A) and $\mathscr{S}_B$ a species partition satisfying (B), then

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

We also observe that for each $\mathcal{X}$-partition $\mathscr{S}$,

$$\mathscr{S}_0 \;\leq\; \mathscr{S} \;\leq\; \mathscr{S}_1,$$

so it does make sense to say that the partition $\mathscr{S}_1$ is the 'coarsest' $\mathcal{X}$-partition and that $\mathscr{S}_0$ is the 'finest' $\mathcal{X}$-partition.

Since we also know that $\mathscr{S}_1$ satisfies (AE), and that $\mathscr{S}_0$ satisfies (BE), we would like to know if there are finer partitions satisfying (AE) and coarser partitions satisfying (BE). However, it makes no sense in general to speak of the coarsest or finest partition satisfying a given property, since the coarsest or finest partition of a given set of partitions may not belong to this set. In the next section, we thus investigate the possibility of defining 'the finest partition satisfying (AE)' as well as 'the coarsest partition satisfying (BE)'.

The lacy and loose species definitions

We state the following result that ensures the existence of the finest partition satisfying (AE) and the coarsest partition satisfying (BE).

Theorem 1. *Given $\mathscr{P}$ and $\mathscr{H}$, there exists a unique finest partition of $\mathcal{X}$ satisfying the exclusivity and heterotypy between species properties (AE), and a unique coarsest partition of $\mathcal{X}$ satisfying the exclusivity and homotypy within species properties (BE).*

This result is proved in SI. It allows us to highlight and name two new different species definitions, each of which satisfies exclusivity and another desirable property. These definitions are illustrated in Figure 4.

Loose species definition. The loose species partition is the finest partition satisfying (AE).

Lacy species definition. The lacy species partition is the coarsest partition satisfying (BE).

For any species partition $\mathscr{S}$ satisfying (E), 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 (or equivalently its associated hierarchy $\mathscr{H}$). For every species S, since S is exclusive there is a unique internal node $u(S)$ of T such that S is exactly constituted of the labels of the tips subtended by $u(S)$. Such a node will hereafter be called a *phylogenetic node*. Then $T_{\mathscr{S}}$ is obtained from T by merging the subtree descending from every phylogenetic node into a single edge. The hierarchy $\mathscr{H}_{\mathscr{S}}$ corresponding to $T_{\mathscr{S}}$ can then be defined in terms of $\mathscr{S}$ and $\mathscr{H}$ 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 4 shows both the lacy phylogeny and the loose phylogeny associated with a simple genealogy and a simple phenotypic partition.

For larger trees, we now describe a simple procedure to get the phylogeny corresponding either to the lacy or to the loose definition, based on the knowledge of the collection $\mathscr{H}$ of genealogical clades and of the phenotypic partition $\mathscr{P}$. Interestingly, building $\mathscr{H}_{\mathscr{S}}$ also offers a quick way to get $\mathscr{S}$ under the lacy and loose definitions, because species are the smallest sets of

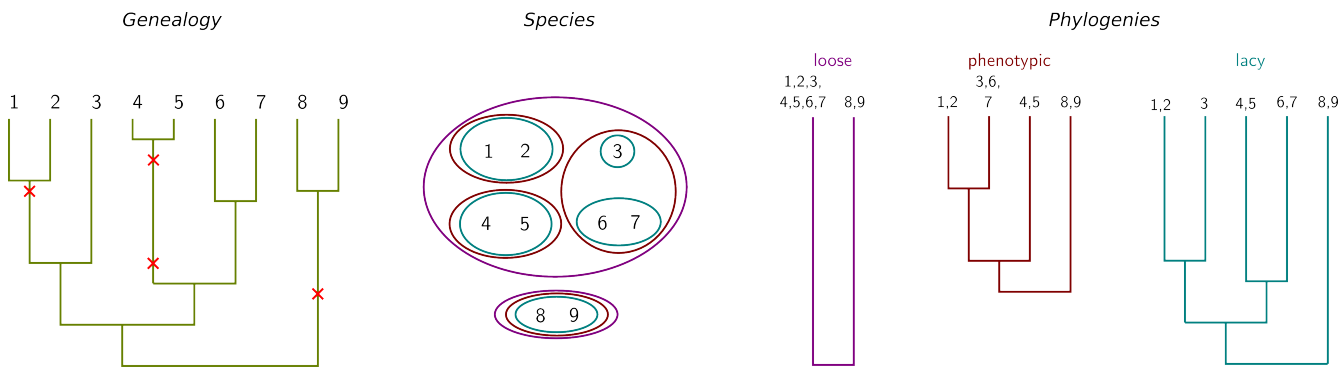

FIGURE 4: Species partitions associated to each of three definitions. Left panel: The fixed genealogy with point mutations (infinite-allele model) leading to the phenotypic partition $\mathscr{P} = \{\{1,2\},\{3,6,7\},\{4,5\},\{8,9\}\}$. Middle panel: Inclusion relations between the three species partitions, 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 phylogeny, phenotypic phylogeny (under the arbitrary convention that divergence times are taken as mutation times), lacy phylogeny.

labels in $\mathscr{H}_{\mathscr{G}}$. The different steps of the algorithm are explained hereafter, illustrated in Figure 5 and formalized in SI.

First, we classify all interior nodes of the genealogy as 'convergent node' or 'divergent node'. An interior node is convergent if there are 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 paraphyletic. Second, we build a phylogeny by deciding which interior nodes are 'phylogenetic nodes', that is, appear in the phylogeny.

Observation 4. *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 ancestors of divergent nodes. Other nodes are declared non-phylogenetic.

Let us say a few words why the last observation holds.

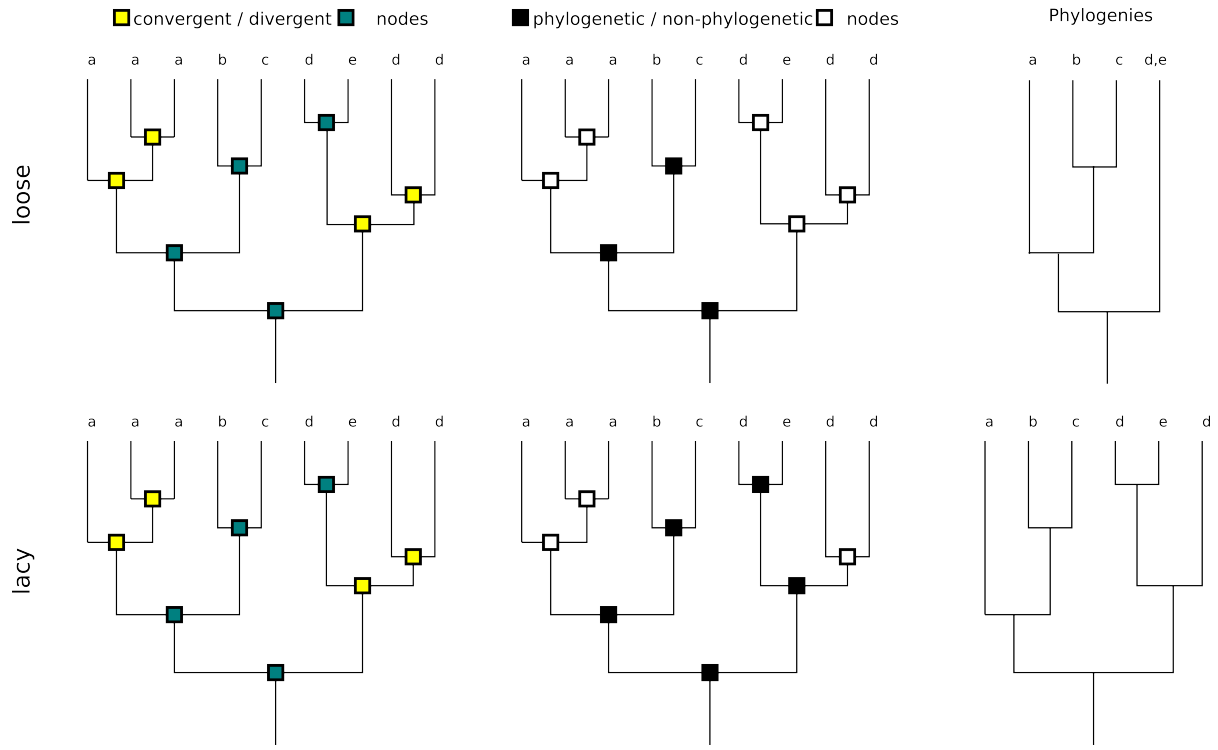

FIGURE 5: Construction of the phylogeny under the lacy and loose species definitions. At the tips, letters correspond to different phenotypes. Left panel: The genealogy with interior nodes classified as convergent (yellow) or divergent (blue). Middle panel: The genealogy with interior nodes classified as non-phylogenetic (white) or phylogenetic (black). Right panel: The corresponding phylogeny. First row, loose: Yellow nodes and blue nodes descending from yellow nodes are colored white. Second row, lacy: Blue nodes and yellow nodes ancestors of a blue node are colored black. Building the loose phylogeny consists intuitively in conserving all nodes, from the root to the tips, until reaching the first yellow node. In contrast, building the lacy phylogeny intuitively consists in discarding all nodes, from the tips to the root, until reaching the first blue node.

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). It is intuitive that 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), 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). It is intuitive that 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 representations attempting to describe evolutionary trees on various scales simultaneously. The multi-species coalescent model is one of the most influential of these representations (Maddison 1997; Degnan and Rosenberg 2009). In this model, a species tree is first specified (e.g., sampled from a given prior distribution on trees) and given the species tree, the gene genealogies are drawn from a censored coalescent (i.e., lineages can coalesce only if they lie in the same ancestral species). In particular, Hudson and Coyne (2002); Rosenberg (2007); Mehta et al. (2016) used this model to assess the relevance of the reciprocal monophyly criteria to recognize species. 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, our approach consists in assuming that macro-evolutionary patterns are

shaped by micro-evolutionary processes. This bottom-up approach has been adopted in other previous studies as well. Let us point out in particular the works of Aldous et al. (2008, 2011) similar in spirit to ours, which make several proposals to lump together lower-order taxa in order to build trees on higher-order taxa. For example, Aldous et al. (2008) propose three different ways of grouping together species into so-called genera. They ground their definitions on the knowledge of a phylogeny with point mutations, with the main difference that each split distinguish a mother and a daughter lineage. All their propositions lead to homotypy within genera, but none of them leads to exclusive genera with respect to the species phylogeny.

To the contrary, we pointed out here three desirable biological properties, among which (E) (exclusivity) is of central importance. We explicitly defined and compared three natural species definitions based on the knowledge of the genealogy and of a phenotypic partition. Each of these satisfies a different set of properties, summarized in Table 1.

Species definition	A	B	E
Lacy		×	×
Loose	×		×
Phenotypic	×	×	

Table 1: Properties of species partitions under different definitions. A: Heterotypy between species; B: Homotypy within species; E: Exclusivity.

Additionally, we showed that no species definition generally satisfies the three desirable properties, unless the phenotypic partition is exlusive in the first place. In particular, we emphasize that the most popular ways of defining species in recent individual-based modeling studies of macro-evolution (i.e., modeling diversification with phenotypic point mutations followed by defining species on a phenotypic basis) lead in general to species partitions that are paraphyletic with respect to the genealogy. In contrast, the loose species definition, previously used in the context of diversification with point mutations (Manceau et al. 2015) systematically yields exclusive species. We extended here this study and compared it to a third species definition also leading to exclusive species partitions, 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.

For systematists, the species problem is tightly linked to the practical problem of clustering individuals on a phenotypic basis. Classifying diversity is notoriously difficult for many reasons, including the difficulty of choosing the appropriate level of description (e.g., morphologic, genetic, behavioral...), the ubiquitous presence of convergent evolution and reversal events, and the difficulty to agree on a unique species concept (see Mayden 1997; Groves 2007; De Queiroz 2007; Baum 2009 for recent discussions around the different species concepts that have been proposed).

On the other hand, one could hope that modelers dealing with abstractions should not encounter the same difficulties in defining a proper species concept. To the contrary, even within the same extremely simplistic framework considering asexual lineages accumulating neutral phenotypic changes through time, several different species definitions have been defined (some of them reviewed in the second section of the paper). All of them fit the general species definition from De Queiroz (2007) of 'separately evolving metapopulation lineages', while satisfying distinct desirable properties. We focused in particular on (A) heterotypy between, and (B) homotypy within species, that are reminiscent of the typological species concepts (Regan 1925; Sneath 1976) as well as on (E) the exclusivity property, that is reminiscent of the genealogical species concept (Avise and Ball 1990) or the 'species as taxa' view (Baum 2009). We argue that introducing different properties and comparing species definitions based on the properties they fulfill in simple models might help shed light on the species problem.

The lacy and loose species definitions may also draw the attention of modelers on the issue of the temporal extent of species, which is an extension of the long-lasting debate around the species definition in evolutionary biology. Species are said to be *synchronic* when the definition is instantaneous. Synchronic supporters usually warn against the use and abuse of the *fossil species* and *species age* concepts, the reason being that a succession of individuals may have changed fundamentally through time, even if some morphological traits are well conserved. From this point of view, the phylogeny does not represent genealogical relations between species, but rather genealogical relationships between contemporary groups of individuals currently

known as species. To the contrary, a number of macroevolutionary studies assume that species are *diachronic*, i.e., they have a birth time and a death time, which correspond respectively to nodes and tips in the phylogeny. In other words, we should be able to decide whether two individuals living at different times belong or not to a same species. Among the three species definitions studied in this paper, only the phenotypic definition is diachronic. Indeed, the partition $\mathscr{S} = \mathscr{P}$ only depends on phenotypes and not on the genealogical history of individuals. As such, we could say that two individuals living at two different times belong to the same species, solely based on the fact that they display the same phenotype. Under the lacy and loose species definitions, the partition $\mathscr{S}$ does depend on the phenotypic partition $\mathscr{P}$ but also on the whole genealogical history up to the present. As a consequence, these definitions are synchronic.

In this work, we have only been interested in defining what species could be in a rather simple modeling framework, and we have refrained from making a stand on how species should be defined in nature. We feel philosophically closer to what has been called the *cynical species concept* (Kitcher 1984), that is, the claim that in fine, species are 'whatever a competent taxonomist chooses to call a species'. Species could be defined differently depending on the taxon considered or on the period of description, and the three species definitions studied in this paper despite their simplicity, could be applied more or less appropriately to model various groups. We propose guidelines to approach real datasets depending on what seems to be more reasonable in each particular clade.

Species described long ago are more likely to be based on phenotypic information alone, and seem thus closer to what we called the phenotypic species partition. On the opposite, the recent rise of molecular methods in evolutionary biology may have brought (E) the exclusivity property to the forefront. The reconstruction of gene genealogies has stimulated the development of methods aiming at automatically delineating putative species from sequence data (Yang and Rannala 2010; Puillandre et al. 2012; Fujisawa and Barraclough 2013; Zhang et al. 2013). Recent species descriptions are thus more likely to concern exclusive groups of individuals than earlier. Further, the choice between the lacy or loose species definition could depend on the 'standard philosophy' of taxonomists in a particular area of the tree of life. Our guess would be that the

lacy definition could be used as an approximation of recent taxonomist work on emblematic clades, where more or less homotypic groups of individuals can be separated on a genealogical basis into what are known as *cryptic species* (see Bickford et al. 2007 for a review and examples of cryptic species). In other clades, taxonomists may prefer to ensure that species are diagnosable and exclusive units, two properties stressed as 'priority taxon naming criteria' by Vences et al. (2013). The loose definition does incorporate genealogical information in order to define the species partition in the first place but in this definition species are composed of distinct phenotypes, so that any observed individual can indeed be assigned to its species on a phenotypic basis.

Note that this confrontation of the theoretical model to the biological reality is far too caricatural. First, we assumed here that molecular characters provide direct access to the true underlying genealogy of individuals, whereas we should in fact treat them as any other phenotypic character. Second, we chose to represent genealogies as trees. This simplification may break down for genealogies of sexually reproducing organisms, for which the question of grouping individuals into taxa is far more complex (Hudson and Coyne 2002; Samadi and Barberousse 2006). Advanced theoretical work in this direction has been undertaken by Dress et al. (2010); Kwok (2011); Alexander (2013); Alexander et al. (2015), in a framework closer to biological reality, but much less connected to most modeling studies in macro-evolution.

Conclusion 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 unit of macro-evolution, are formed and deformed by these processes remains a major challenge. We hope that presenting and comparing explicitly new species definitions in a simple framework will help make a step forward toward a better integration of the individual level into models of diversification.

Acknowledgements The authors are very grateful to R.S. Etienne and M. Steel for their comments on previous drafts of this paper, and to David Baum for helpful litterature advice. The authors thank the Center for Interdisciplinary Research in Biology (CIRB, Collège de France) for funding.

References

Aguilée, R., D. Claessen, and A. Lambert. 2013. Adaptive radiation driven by the interplay of eco-evolutionary and landscape dynamics. Evolution 67:1291–1306.

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.

Barraclough, T. G. and S. Nee. 2001. Phylogenetics and speciation. Trends Ecol. Evol. 16:391–399.

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.

Chave, J. 2004. Neutral theory and community ecology. Ecol. Lett. 7:241–253.

Crow, J. F. and M. Kimura. 1970. An introduction to population genetics theory. Harper & Row, New York.

Davies, T. J., A. P. Allen, L. Borda-de Água, J. Regetz, and C. J. Melián. 2011. Neutral biodiversity theory can explain the imbalance of phylogenetic trees but not the tempo of their diversification. Evolution 65:1841–1850.

De Aguiar, M. A. M., M. Baranger, E. M. Baptestini, L. Kaufman, and Y. Bar-Yam. 2009. Global patterns of speciation and diversity. Nature 460:384–387.

De Queiroz, K. 2005. Ernst Mayr and the modern concept of species. Proc. Natl. Acad. Sci. U. S. A. 102:6600–6607.

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.

Degnan, J. H. and N. A. Rosenberg. 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends Ecol. Evol. 24:332–340.

Dieckmann, U. and M. Doebeli. 1999. On the origin of species by sympatric speciation. Nature 400:354–357.

Doebeli, M. and U. Dieckmann. 2003. Speciation along environmental gradients. Nature 421:259–264.

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.

Estes, S. and S. J. Arnold. 2007. Resolving the paradox of stasis: models with stabilizing selection explain evolutionary divergence on all timescales. Am. Nat. 169:227–244.

Etienne, R. S. and B. Haegeman. 2011. The neutral theory of biodiversity with random fission speciation. Theor. Ecol. 4:87–109.

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

Etienne, R. S. and J. Rosindell. 2011. Prolonging the past counteracts the pull of the present: Protracted speciation can explain observed slowdowns in diversification. Syst. Biol. 61:204–213.

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.

Geritz, S. A., J. A. Metz, É. Kisdi, and G. Meszéna. 1997. Dynamics of adaptation and evolutionary branching. Phys. Rev. Lett. 78:2024–2027.

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.

Groves, C. 2007. Species concepts and speciation: Facts and fantasies. Pages 1861–1879 *in* Handbook of Paleoanthropology (W. Henke and I. Tattersall, eds.). Springer Berlin Heidelberg.

Hartl, D. L. and A. G. Clark. 1997. Principles of population genetics vol. 116. Sinauer associates Sunderland.

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.

Kingman, J. F. C. 1982. The coalescent. Stoch. Proc. Appl. 13:235–248.

Kitcher, P. 1984. Species. Philos. Sci. 51:308–333.

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.

Latimer, A. M., J. A. Silander, and R. M. Cowling. 2005. Neutral ecological theory reveals isolation and rapid speciation in a biodiversity hot spot. Science 309:1722–1725.

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

Mallet, J. 2001. Species, concepts of. Enc. Biodivers. 5:427–440.

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.

Metz, J. A., S. A. Geritz, G. Meszéna, F. J. Jacobs, and J. S. Van Heerwaarden. 1996. Adaptive dynamics, a geometrical study of the consequences of nearly faithful reproduction. Pages 183–231 *in* Stochastic and Spatial Structures of Dynamical Systems (S. J. van Strien and S. M. Verduyn Lunel, eds.).

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.

Nei, M., T. Maruyama, and C.-I. Wu. 1983. Models of evolution of reproductive isolation. Genetics 103:557–579.

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.

Pigot, A. L., A. B. Phillimore, I. P. F. Owens, and C. D. L. Orme. 2010. The shape and temporal dynamics of phylogenetic trees arising from geographic speciation. Syst. Biol. 59:660–673.

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.

Rosenberg, N. A. 2007. Statistical tests for taxonomic distinctiveness from observations of monophyly. Evolution 61:317–323.

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.

Velasco, J. D. 2009. When monophyly is not enough: exclusivity as the key to defining a phylogenetic species concept. Biol. Philos. 24:473–486.

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.

Zhang, J., P. Kapli, P. Pavlidis, and A. Stamatakis. 2013. A general species delimitation method with applications to phylogenetic placements. Bioinformatics 29:2869–2876.

List of Figures

1 Different scenarios leading to paraphyletic phenotypic partitions, for the 'red' phenotype. Left panel, convergence: The same phenotype arises twice independently in two branches. Middle panel, reversal: A new phenotype arises, and disappears later. Right panel, ancestral type retention: The group of individuals showing the ancestral phenotype is not exclusive. 7

2 The five modes of speciation proposed in individual-based models of macro-evolution. In each panel, the genealogy (green tree) of individuals (integer labels) is given on the left, along with mutations (red crosses) that confer new types (greek letters) to individuals (infinite-allele model). The corresponding species partition is represented on the right of each panel (subsets of labels circled in red). a) *Speciation by point mutation.* Two individuals are in the same species if and only if they carry the same type. b) *Speciation by random fission.* The same mutation hits several individuals simultaneously and confers the same new type to their descents. Two individuals are in the same species if and only if they carry the same type. c) *Protracted speciation.* Two individuals are in the same species if their divergence/coalescence time is smaller than a given threshold (grey dashed line) or if they carry the same type. d) *Speciation by genetic incompatibility.* Two individuals are said compatible if there are less than q mutations they do not share; species are the connected components of the compatibility graph. e) *Speciation by genetic differentiation.* Species are the smallest exclusive groups of individuals, such that individuals carrying the same type always belong to the same group. 10

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: The 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. . 13

4 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 phylogeny, phenotypic phylogeny (under the arbitrary convention that divergence times are taken as mutation times), lacy phylogeny. 18

5 Construction of the phylogeny under the lacy and loose species definitions. At the tips, letters correspond to different phenotypes. Left panel: The genealogy with interior nodes classified as convergent (yellow) or divergent (blue). Middle panel: The genealogy with interior nodes classified as non-phylogenetic (white) or phylogenetic (black). Right panel: The corresponding phylogeny. First row, loose: Yellow nodes and blue nodes descending from yellow nodes are colored white. Second row, lacy: Blue nodes and yellow nodes ancestors of a blue node are colored black. Building the loose phylogeny consists intuitively in conserving all nodes, from the root to the tips, until reaching the first yellow node. In contrast, building the lacy phylogeny intuitively consists in discarding all nodes, from the tips to the root, until reaching the first blue node. 19

Supplementary Information
The species problem from the modeler's point of view

Marc Manceau[1,3,4], Amaury Lambert[2,3]

[1]*Muséum National d'Histoire Naturelle, 75005 Paris, France;*

[2]*Laboratoire Probabilités et Modèles Aléatoires, UPMC Univ Paris 06, 75005 Paris, France;*

[3]*Center for Interdisciplinary Research in Biology, Collège de France, CNRS UMR 7241, 75005 Paris, France;*

[4]*Institut de Biologie, École Normale Supérieure, CNRS UMR 8197, 75005 Paris, France*

Some of the results stated in Sections A and B are classical results in combinatorics for partially ordered sets (Bóna 2011). 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.

- 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

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

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

$$\Sigma_E := \{\mathcal{X}\text{-partitions satisfying (E)}\}$$

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

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

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

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

$$\mathscr{S} \in \Sigma_E \Longleftrightarrow \mathscr{S} \subseteq \mathscr{H}.$$

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

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

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

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

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

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

The proof is divided into two parts. First, given a set of partitions Σ (resp. $\Sigma \subseteq \Sigma_E$), 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 $\inf \Sigma_{AE} \in \Sigma_{AE}$ and $\sup \Sigma_{BE} \in \Sigma_{BE}$, hence yielding the definitions $\mathscr{S}_{\text{loose}} := \inf \Sigma_{AE}$ and $\mathscr{S}_{\text{lacy}} := \sup \Sigma_{BE}$.

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 $\Sigma \subseteq \Sigma_E$, $\overline{\mathcal{R}}_\Sigma$ 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 $\Sigma \subseteq \Sigma_E$ 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 $\Sigma \subseteq \Sigma_E$, $\mathscr{S}_1 \subseteq \mathscr{H}$ 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. $\square$

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 $\Sigma \subseteq \Sigma_E$, 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, 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 $\Sigma \subseteq \Sigma_E$. 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 $\Sigma \subseteq \Sigma_E$ 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 by $\overline{\mathcal{R}}_\Sigma$ and by definition of $\overline{\mathcal{R}}_\Sigma$, either $x \notin S'$ or $y \notin S'$ and we got the contradiction. $\square$

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 $\Sigma \subseteq \Sigma_E$, 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 $\inf \Sigma_{AE} \in \Sigma_{AE}$ *and* $\sup \Sigma_{BE} \in \Sigma_{BE}$

In order to prove that $\inf \Sigma_{AE} \in \Sigma_{AE}$ and $\sup \Sigma_{BE} \in \Sigma_{BE}$, 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 $\Sigma \subseteq \Sigma_E$, for any $S \in \sup \Sigma$, S can be written in the form of the following non-empty union

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

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 $S = S'$. First recall thanks 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 $\Sigma \subseteq \Sigma_E$ and let $S \in \sup \Sigma$. Set

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

and let us prove that $S = S'$. First recall thanks 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 thanks 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 $F \subseteq S_1 \subseteq S$. 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 $\mathscr{S}_1, \mathscr{S}_2 \in \Sigma \subseteq \Sigma_E$, 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 $\{x, y\} \subseteq S^* \subset S$, which shows that $F = F_1 \cup \{x\} \subseteq S^* \subseteq S$ and terminates the proof. $\square$

We can now turn to the end of the proof of Theorem 1.

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

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

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

(ii) $\inf \Sigma_{AE} \in \Sigma_E$**.** Consider $S \in \inf \Sigma_{AE}$. From Lemma 3, we get

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

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

(iii) $\sup \Sigma_{BE} \in \Sigma_B$**.** Consider $S \in \sup \Sigma_{BE}$ and recall from Lemma 3 that there is $\mathscr{S} \in \Sigma_{BE}$ 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 $\sup \Sigma_{BE} \in \Sigma_B$.

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

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 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

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

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

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

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