bioRxiv preprint doi: https://doi.org/10.1101/2023.11.09.564915; this version posted November 10, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Opening the species box: What microscopic models of neutral speciation have to say about macroevolution

Élisa Couvert[1,2,3] * François Bienvenu[4] Jean-Jil Duchamps[5] Adélie Erard[3]
Verónica Miró Pina[6] Emmanuel Schertzer[7] Amaury Lambert[1,2]

[1] *Institut de Biologie de l'ENS* (IBENS), École Normale Supérieure, PSL Université, CNRS UMR8197, INSERM U1024, Paris, France

[2] *Centre Interdisciplinaire de Recherche en Biologie* (CIRB), Collège de France, PSL Université, CNRS UMR7241, INSERM U1050, Paris, France

[3] *Université Paris Cité*, CNRS UMR8145, MAP5, F-75006, Paris, France

[4] *Institute for Theoretical Studies*, ETH Zürich, 8092 Zürich, Switzerland

[5] *Université de Franche-Comté*, CNRS, LmB, F-25000 Besançon, France

[6] *Centre for Genomic Regulation* (CRG), The Barcelona Institute of Science and Technology, Barcelona, Spain; Universitat Pompeu Fabra (UPF), Barcelona, Spain

[7] *Faculty of Mathematics*, University of Vienna, Oskar-Morgenstern-Platz 1, 1090 Wien, Austria

Abstract

In the last two decades, lineage-based models of diversification, where species are viewed as particles that can divide (speciate) or die (become extinct) at rates depending on some evolving trait, have been very popular tools to study macroevolutionary processes. Here, we argue that this approach cannot be used to break down the inner workings of species diversification and that "opening the species box" is necessary to understand the causes of macroevolution.

We set up a general framework for individual-based models of neutral speciation (i.e. no selection forces other than those acting against hybrids) that rely on a minimal number of mechanistic principles: (i) reproductive isolation is caused by excessive dissimilarity between pheno/genotypes; (ii) dissimilarity results from a balance between differentiation processes and homogenization processes; and (iii) dissimilarity can feed back on these processes by decelerating homogenization.

We classify such models according to the main process responsible for homogenization: (1) clonal evolution models (ecological drift), (2) models of genetic isolation (gene flow) and (3) models of isolation by distance (spatial drift). We review these models and their specific predictions on macroscopic variables such as species abundances, speciation rates, interfertility relationships, phylogenetic tree structure...

We propose new avenues of research by displaying conceptual questions remaining to be solved and new models to address them: the failure of speciation at secondary contact, the feedback of dissimilarity on homogenization, the emergence in space of reproductive barriers.

*elisa.couvert@college-de-france.fr

bioRxiv preprint doi: https://doi.org/10.1101/2023.11.09.564915; this version posted November 10, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

1 Introduction

1.1 Phylogenetic approaches to diversification: let us open the species box

Starting with the seminal works of the paleontologists of the "Woods Hole group" (Raup et al., 1973) and drawing on parallel mathematical progress (Kendall, 1948; Nee et al., 1994; Aldous, 2001; Aldous and Popovic, 2005), a powerful quantitative method has been developed in macroevolution, using birth-death processes as models for the diversification of species. In these so-called *lineage-based models of diversification*, species are particles that can undergo two kinds of events: speciation, modeled by instantaneous division; and extinction, modeled by instantaneous death. The phylogenetic patterns (as quantified by balance indices and other shape statistics, see Box 4) predicted by lineage-based models can then be studied mathematically and used either to test whether a birth-death process, seen as a null model, can explain the observed phylogeny; or, alternatively, to infer how speciation and extinction rates may depend on some evolving trait carried by the species. This so-called *phylogenetic approach to diversification* has been very popular in macroevolution and surveyed multiple times in the last decade (Ricklefs, 2007; Pyron and Burbrink, 2013; Stadler, 2013; Morlon, 2014).

However, when it comes to processes as complex – and occurring on time scales as long – as the diversification of species, this approach suffers from several limitations:

- The build-up of genetic differentiation between populations that leads to the formation of new species takes time, and so do the demographic declines and population extirpations that lead to species extinctions. This is not captured by coarse-grained lineage-based models where speciation and extinction are instantaneous;

- A mere phylogeny contains little signal, which gives rise to statistical problems like false associations between rate and trait (Rabosky and Goldberg, 2015) or non-identifiability of parameters (Louca and Pennell, 2020) – but see also Morlon et al. (2022) for a discussion;

- Phylogenies are nowadays built by comparing sequences of nucleic acids, hence the name "molecular phylogenies". However, because of recombination and, as increasingly recognized, gene flow between species (Marques et al., 2019; Pennisi, 2016), different genes can have very different genealogies, so that a phylogeny is a rather poor summary of evolutionary history (Degnan and Rosenberg, 2006; Maddison, 1997);

- Speciation and extinction rates are useful notions to understand the diversification process but, *in fine*, are only compounded quantities that summarize very crudely some fine-scale phenomena such as: habitat selection, species sorting, divergent adaptation, reproductive isolation, assortative mating, introgression, reinforcement, speciation collapse, etc. If we want to characterize the processes that generate biodiversity, we have to understand and infer these processes (Li et al., 2018). Actually, the way these rates and other macroevolutionary or macroecological observables depend on these fine-scale processes remains one of the most intriguing questions in macroevolution.

To address these questions, an alternative way to study diversification processes consists in moving away from the assumption that species are particles (top-down approach) to directly model how the elementary constituents of species (populations, individuals, genomes) all work together to lead to speciation (bottom-up approach).

bioRxiv preprint doi: https://doi.org/10.1101/2023.11.09.564915; this version posted November 10, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

1.2 A plea for a bottom-up approach to speciation

We call models that are rooted in the fine-scale description of ecological and genetic phenomena *microscopic models*. Building and using these models is a 4-step process drawing on complementary tools and areas of expertise.

First, microscopic models are specified by relying on **mechanistic principles** and **measurable parameters** (e.g., dispersal rate, demographic parameters, mutation rate). These models can be of two types: parameter-rich models, whose precision and realism sacrifice simplicity and which can easily suffer from over-fitting flaws; and parsimonious models. We use the term *archetypal* to denote models that are both **microscopic** and **parsimonious**. From now on, we exclusively focus on such archetypal models of speciation.

As a second step, **mathematical micro–macro approaches** can provide a description of large-scale phenomena in terms of the microscopic parameters, sometimes at the cost of taking the parameters to some extremal region of the parameter space (e.g., unlimited dispersal, also called mean field limit; large community size; small mutation rate; etc.).

The third step then consists in deriving accurate predictions of the model for some **macroscopic observables** corresponding to biological quantities of interest, either at the level of the species (intraspecific genetic diversity, species abundance, species range) or at the level of communities (species richness, speciation/extinction rate, phylogeny). These macroscopic observables can also be more complex patterns, such as relations between two kinds of observables (species-area relationship, species-speciation rate relationship, ecosystem functioning-diversity relationship, gene tree-species tree coupling).

The last step consists in confronting the model to reality. This involves tuning the parameters to see how well the predictions of the model can fit the data (typically, using maximum likelihood estimates).

When successful – which requires that the mathematical analysis work and that the predictions fit the data –, this methodology produces models that

- are biologically realistic (step 1);
- are universal, in the sense that they reflect the general behavior of a large class of more detailed models (step 2);
- yield predictions that can readily be linked back to the basic assumptions (step 3);
- have few parameters and therefore are easily falsifiable, which gives more confidence in their explanatory power (step 4).

1.3 The species definition problem

The bottom-up approach to speciation, in addition to its aforementioned concrete advantages, also gives insight into what makes a species a species. Indeed, despite being central to evolutionary biology, the notion of species is not well-defined, in the sense that there is no single agreed formal definition. One of the most widely accepted ideas is that species should correspond to "groups of reproductively compatible populations that are reproductively incompatible with other such groups". This is the so-called *biological species concept*, or BSC; see Coyne et al. (2000) for a detailed discussion.

However, despite its apparent simplicity – and considerable influence in biology – the BSC admits more than one mathematical formalization (see Box 1). More than a problem with the BSC, this reflects the fact that species are complex entities with fuzzy boundaries. Much like when

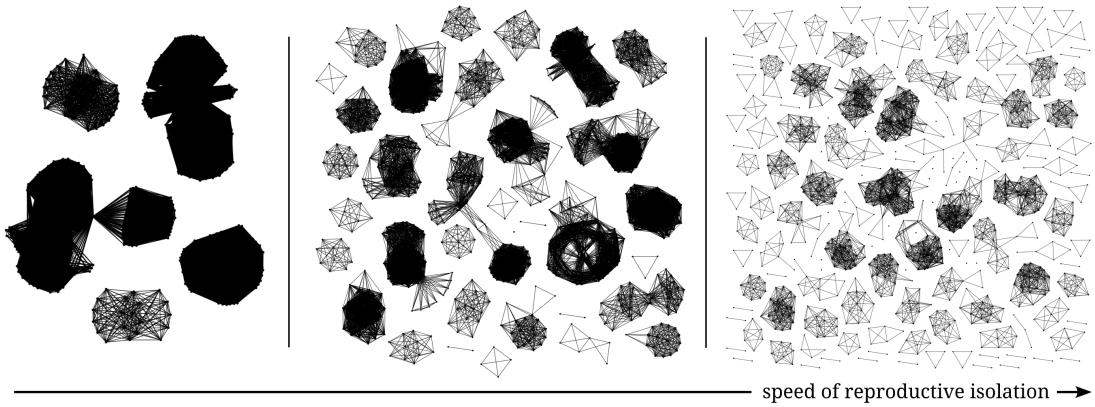
bioRxiv preprint doi: https://doi.org/10.1101/2023.11.09.564915; this version posted November 10, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

trying to define a heap of sand, it may not be possible to find a one-size-fits-all definition. But this does not prevent dunes from existing, nor does it make it impossible to study their formation and dynamics. Only by opening the species box and by choosing the right level of description to model the elementary constituents of species – metapopulations, populations, individuals, genes – can we hope to circumvent the "species problem" and to model the formation of species without being tied to a specific (and inevitably imperfect) definition.

Box 1: **The biological species concept and the interfertility graph.**

The *biological species concept* (BSC) considers species to be "groups of reproductively compatible populations that are reproductively incompatible with other such groups". If this were a strict definition, then phenomena such as ring species (where populations of the same species are reproductively incompatible, see Irwin et al. 2001) and hybrid speciation (where populations of different species hybridize to produce a new species, see Mallet 2007) would be impossible. Therefore, the BSC tells us what the essence of an idealized species should be, but it is not an accurate description of species in the real world.

The BSC framework postulates that the key criterion on which species should be defined is the notion of *interfertility*, i.e. the ability to interbreed. Interfertility is, ultimately, a relation between individuals; but as a first step in opening the species box it can be seen as a relation between populations. This relation can be represented by a graph whose vertices correspond to populations, and where two vertices are linked by an edge if and only if the two corresponding populations are reproductively compatible: we call this graph the *interfertility graph*. In an ideal setting, the interfertility graph would be a disjoint union of cliques (i.e. groups of vertices such that there is an edge between every pair of vertices), but – as we have seen – it will deviate from this ideal in practice. Bienvenu et al. (2019) introduced a mathematical model to study the structure and dynamics of interfertility graphs. The following graphs are simulations from that model, for 1000 populations and various values of its drift parameter.



Given their interfertility graph, there is a continuum of ways to partition a set of populations into species, as illustrated in Figure 1. At one end of the spectrum, if we want to ensure that reproductively compatible populations belong to the same species, then species have to be defined as maximal cliques; however, some drawbacks of this approach are that there may not be a unique partition into maximal cliques, and that it may allow too much hybridization. At the other end of the spectrum, species can be defined as connected components of the interfertility graph. This definition has the advantage of being unambiguous and very natural, in that two populations belong to the same species if and only if there is some possibility of (direct or indirect) gene flow between them; however, it can lead to species being composed mostly of reproductively incompatible populations, and it does not allow any hybridization. Thus, there is no "one-size-fits-all" definition, and which definition proves most relevant may depend on the specific setting.

bioRxiv preprint doi: https://doi.org/10.1101/2023.11.09.564915; this version posted November 10, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

On the other hand, for any microscopic model representing the speciation process, one has to decide after letting the process unfold which groups of populations or individuals have to be considered species, in order to uncover the macroecological predictions of the model. This procedure, which consists in assigning to a species each individual/population of a sample, is known as the *species clustering problem*. This problem is not only relevant to our theoretical understanding of speciation: it also arises as a very concrete question in bacterial genomics, particularly in the context of barcoding and metabarcoding.

In view of the BSC, a natural way to group individuals/populations into species is to consider all relations of interfertility and to partition the associated interfertility graph into groups such that each pair of elements in the same group are linked by a chain of interbreeding elements, see Box 1. Figure 1 displays two such partitions: into maximal cliques or into connected components.

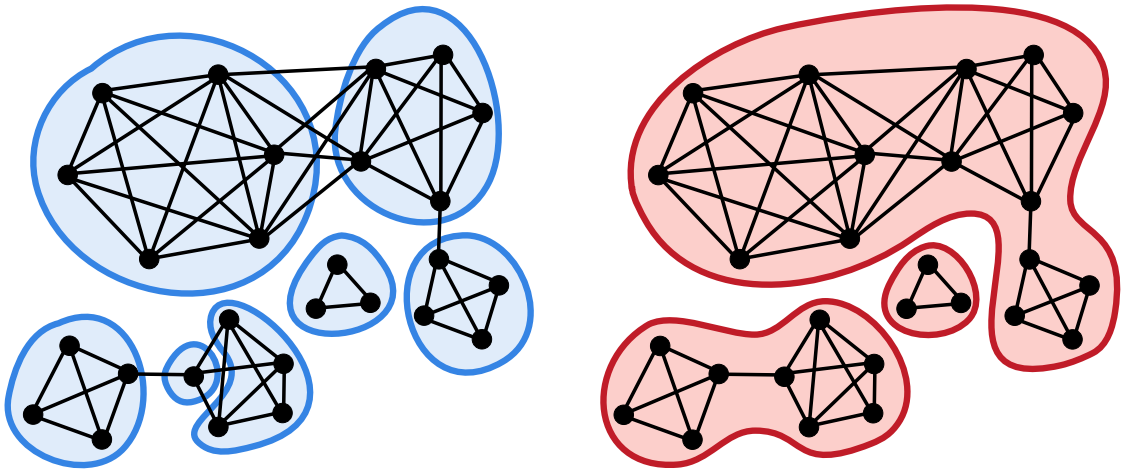


Figure 1: Several ways to partition the vertices of a graph. On the left, in blue, a partition into maximal cliques. On the right, in red, the partition into connected components.

However in most cases, the interfertility graph is not available and one has to resort to other kinds of data – typically genetic distances. Given a matrix of genetic distances between individuals, the following algorithm, which we refer to as the *threshold clustering algorithm*, provides a solution to the species clustering problem:

- Fix a threshold distance;
- Consider the graph where any two individuals are linked if and only if the genetic distance between them is smaller than this threshold;
- Species are then defined to be the connected components in this graph.

For example, metabarcoding studies traditionally use a threshold clustering algorithm to delimit species (called "operational taxonomic units" or OTUs in that context), using the percentage of pairwise differences on 16S ribosomal RNA for the genetic distance, and a threshold equal to $\delta = 0.03$.

Many microscopic models of speciation implicitly use a threshold clustering algorithm to define species. In most cases, lineages are assumed to accumulate differences under the *infinite-allele model* (Kimura and Crow, 1964). In this setting, each mutational event gives rise to a new allele that has never existed in the past and increases the genetic distance to each ancestor by 1. More formally, the infinite-allele assumption endows the alleles with a tree structure called the *allelic tree*, see Figure 2: allele A is the mother of allele B if the mutation giving rise to allele B occurred

bioRxiv preprint doi: https://doi.org/10.1101/2023.11.09.564915; this version posted November 10, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

on a lineage carrying allele A. The genetic distance between two individuals is merely the graph distance between their alleles in the allelic tree. In the case when the threshold is taken to be 1, the species partition thus obtained is the finest partition such that two individuals carrying the same allele are found in the same species.

Appealing though they may be for their elegance and practicality, a major pitfall of threshold clustering algorithms is that, even under simple models of evolution of genetic distances, they do not always yield species partitions that are compatible with the genealogy – that is, the corresponding "species" are not always monophyletic. This raises the fundamental question of the existence of a natural species partition such that (1) each species forms a subtree of the genealogy and (2) any two individuals that are at genetic distance smaller than the threshold are in the same species? Manceau and Lambert (2019) showed that there is a unique finest species partition satisfying the two conditions above, and gave a simple algorithm to find it.

1.4 Archetypal models of neutral speciation

Throughout the paper, we will restrict our study to neutral speciation, in the sense that it is assumed that no selection is acting during the course of speciation, other than that deriving from hybrid depression. We will thus focus on a class of parsimonious, microscopic models, hereafter referred to as *archetypal models of neutral speciation*, that rely only on the following basic assumptions to represent the speciation process:

1. Speciation is driven by reproductive isolation between individuals or populations, itself resulting from excessive phenotypic dissimilarity. Typically, individuals are endowed with a partially heritable trait (genotype or phenotype) that can undergo evolutionary changes over time. Genetic differentiation between populations, also called population differentiation, builds up as spatially segregating mutations fix in these populations, due to genetic drift, founder effects or divergent adaptation. Dissimilarity can then be quantified as a measure of the distance between trait values, and this distance can be used to cluster individuals into species (see Box 1).

2. Phenotypic dissimilarity is the result of a balance between two processes: spontaneous differentiation driven by mutation as in the previous item, and homogenization processes such as reproduction and migration. For example, during the time when partially isolated populations are in the "grey zone" (no longer a clear species and not yet two reproductively isolated species, De Queiroz, 2007; Roux et al., 2016), old alleles can spread back and replace new alleles, resulting in the failure of these ephemeral populations to speciate (speciation collapse). Within this framework, reproductive isolation can occur when differentiation predominates over homogenization or when historical contingency factors disrupt the equilibrium between these two processes.

3. Dissimilarity can feed back on homogenization processes, by inhibiting them (e.g., outbreeding depression) or promoting them (e.g., disassortative mating). For example, the ability to interbreed may decrease continuously as a function of phenotypic dissimilarity, or it may disappear abruptly when dissimilarity exceeds a certain threshold, e.g., representing genetic incompatibilities (Corbett-Detig et al., 2013; Coyne and Orr, 1989; Matute and Cooper, 2021) (see Box 2).

We categorize these archetypal models into three classes: **(1) clonal evolution models**, **(2) models of genetic isolation**, and **(3) models of isolation by distance**. In all cases,

bioRxiv preprint doi: https://doi.org/10.1101/2023.11.09.564915; this version posted November 10, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

differentiation occurs spontaneously as a consequence of mutation. What differentiates these three classes is the main process responsible for homogenization, namely: **(1) ecological drift** (births, deaths), **(2) gene flow** (sexual reproduction, migration between demes), and **(3) spatial drift** (dispersal). Of course, many models mix these three types of homogenization processes; but most models put the emphasis on one of these barriers to differentiation, with various degrees of purity. Let us be more specific:

1. **Clonal evolution models** assume that differentiation between individuals and populations can only increase as traits/genomes evolve/mutate in parallel. In turn, speciation can only be slowed down by birth and death events purging this diversity. These models include: the point-mutation model of speciation (Hubbell, 2001; Chave, 2004; Jabot and Chave, 2009); the model of protracted speciation (Rosindell et al., 2010; Etienne and Rosindell, 2012; Lambert et al., 2015); and the model of speciation by genetic differentiation (Manceau et al., 2015).

2. **Models of genetic isolation** specifically ask whether individuals or populations are genetically compatible. When this is the case, differentiation can decrease under the effect of a phenomenon tending to homogenize gene pools, e.g. reproduction or gene flow. Furthermore, the rates of homogenization processes may decrease with increasing dissimilarity. These models generally use a more or less faithful representation of genomes (number of loci, physical linkage, etc.), see Box 2. They include the Higgs–Derrida model (Higgs and Derrida, 1991, 1992); the split-and-drift random graph (Bienvenu et al., 2019); the parapatric model of Gavrilets (2000) and its extensions (Yamaguchi and Iwasa, 2013, 2015; Miró Pina and Schertzer, 2019); and the gene-based diversification model (Marin et al., 2020).

3. **Models of isolation by distance** consider spatially embedded populations and seek to study the build-up of species boundaries in space. Homogenization here is mainly controlled by dispersal, range expansion or local extinction. We discard from this category static models of metapopulations composed of a few discrete demes connected by migration. In contrast, we assume here that individuals are positioned in large, regular grids or in continuous space where distance can be physically measured. In these models, species emerge as a result of isolation by distance in addition to possible other mechanisms such as genetic differentiation. These include the "extinction-recolonization" model of Gavrilets et al. (2000a), the "topopatric" model of de Aguiar et al. (2009) and the geographic model of speciation of Pigot et al. (2010).

As explained in the previous section, despite their simplicity archetypal models can make specific predictions on a range of macroecological and macroevolutionary observables that can be confronted to real data, for example:

- Interfertility relationships ("who can interbreed with whom", see Box 1);
- Measures of genomic diversity (see Box 2): genetic diversity within/between species, distribution of genetic differentiation along the genome;
- Measures of species diversity (see Box 3): species richness, species abundance distribution (SAD) – and, in the case of spatial models: range size distribution and spatial distribution of species (species-area relationships (SAR), alpha, beta and gamma diversity);
- Phylogenetic patterns (see Box 4): speciation and extinction rates, phylogenetic balance, lineages-through-time plots, phylogenetic diversity, the shape and coupling of gene trees and species trees.

bioRxiv preprint doi: https://doi.org/10.1101/2023.11.09.564915; this version posted November 10, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

In the next section, we review the three main classes of archetypal models of speciation defined above and their macroscopic predictions. In the last section, we propose some promising avenues of research, conceptual questions to be solved and new models to address them.

Box 2: **The genetic architecture of reproductive isolation.**

In models considering the genetic basis of reproductive isolation (RI), individuals are endowed with a haploid or diploid genome with L loci carrying alleles taking values in a discrete (usually finite) set. These models mainly fall into two categories.

• **RI as a by-product of genetic distance.** We think of the genetic distance d between two genomes g and g' as

$$d(g, g') = \sum_{i=1}^{L} \delta(A_i, A'_i) \tag{1}$$

where A_i (resp. A'_i) is the allele carried at locus i by genome g (resp. g'), and δ is a distance in the allele space. Here, the idea is that excessive genetic distance alters the frequency of mating events (premating isolation) or their outcome (postmating isolation), because the accumulation of differences on a locus-by-locus basis increases the dissimilarity between phenotypes that can directly (e.g., shape of genitalia) or indirectly (e.g., phenology) be related to reproduction.

Nei et al. (1983) used this framework with $L = 1$ or $L = 2$ and a stepwise mutation model, forbidding reproduction when $\delta(A_1, A'_1)$ or $\delta(A_2, A'_2)$ is strictly larger than 1, where $\delta(A, A')$ is the number of mutations needed to go from A to A'. Higgs and Derrida (1991, 1992) introduced a model with two alleles at each locus, forbidding reproduction when $d > d_{\min}$, with $\delta(A, A') = 1/2L$ if $A \neq A'$ (and 0 otherwise).

Genetic differences can arise due to local adaptation to different environments. The "genic view of speciation" (Via, 2009; Wu, 2001) predicts that regions of the genome flanking key ecological loci undergo effectively reduced gene flow. These regions are called genomic islands of differentiation or genomic islands of speciation (Seehausen et al., 2014; Wolf and Ellegren, 2017) and act as barrier loci against gene flow. These barriers are believed to evolve as defenses against outbreeding depression caused by mating with migrants breaking linkage between locally adapted alleles.

• **RI due to genetic incompatibilities.** Genetic incompatibilities are negative epistatic interactions between alleles at *different* loci affecting the success of mating events, altering for example zygote formation, zygote viability (intrinsic isolation) or hybrid fitness (extrinsic isolation).
The starting point of this approach is the so-called **Dobzhansky-Muller incompatibility** (DMI) (Dobhansky, 1937; Muller, 1942) involving two loci. The ancestral genotype is $AABB$. In one subpopulation, a mutation is fixed at the first locus ($aaBB$). In another subpopulation, a mutation is fixed at the second locus ($AAbb$). Hybridization would lead to the co-occurrence of incompatible alleles a and b in the inviable genotype $AaBb$.
Several examples of DMI have been observed in nature. For example, in some marine invertebrates, the two loci correspond to a sperm protein and an egg protein (Vacquier and Swanson, 2011). In the *Xiphophorus* family, they correspond to an oncogene and its repressor: hybrids lack the repressor and develop melanomas due to an excess of melanocyte proliferation (Gordon, 1931; Coyne and Orr, 1989; Patton et al., 2010).

In the case of L loci, the most frequent way to model negative epistatic interactions is by a system of $\binom{L}{2}$ independent locks associated to all pairs of loci supported by two different gametes. Then the reproductive compatibility c between two haploid gametes g and g' can be expressed as

$$c(g, g') = \prod_{i=1}^{L} \prod_{j \neq i} \gamma_{ij}(A_i, A'_j), \tag{2}$$

bioRxiv preprint doi: https://doi.org/10.1101/2023.11.09.564915; this version posted November 10, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

where $\gamma_{ij}(A, A')$ is the probability that the potential reproductive barrier associated to loci i and j is inactive when the first gamete carries allele A at locus i and the second gamete carries allele A' at locus j. Note the contrast of (2) with (1). A common assumption introduced by Orr (Orr (1995); Orr and Orr (1996); Orr (1996)) is that all $\gamma_{ij}(A, A')$ are equal to 1 if A and A' are ancestral alleles, and to some $\gamma < 1$ otherwise. Then the compatibility equals

$$c = \gamma^{\binom{K}{2}},$$

where K is the number of mutated loci in at least one of the two gametes, leading to the so-called "snowball effect": if the number of substitutions increases linearly with time, the number of incompatibilities – and therefore the probability that two individuals cannot interbreed – increases quadratically with time.

The aforementioned models can be understood under the metaphor of "holey" adaptive landscapes suggested by Gavrilets (1997). In this type of fitness landscape, two reproductively incompatible genotypes are connected by a chain of intermediary equally fit genotypes, forming a "ridge". This makes it possible for two reproductively isolated species to emerge without having to cross a fitness valley. This idea can be extended to higher dimensional genotype spaces (Gavrilets and Gravner, 1997), and more complex models have also been proposed; see Gavrilets (2014) for a review.

Box 3: **Macroecological metrics.**

Microscopic models of speciation can offer a better understanding of the mechanisms that underlie commonly observed patterns in macroecology, measured by the following metrics.

• **The Species Abundance Distribution (SAD)** in a given community gives the number of species with abundance n, for all n. A remarkable fact is that the SAD of most communities displays a "hollow" shape (Williams et al., 1964; Magurran, 2003), where a handful of species are abundant and most other species are rare (Hubbell, 2001).

Multiple parametric models have been proposed to fit this hollow curve, such as Fisher's logseries (Fisher et al., 1943a) or Preston's lognormal distribution (Preston, 1948), and many other authors have tried to explain the empirical shape of SADs, either via purely statistical or by mechanistic arguments. See e.g. McGill et al. (2007) for an account of the theories.

• **The Species-Area Relationship (SAR)** gives the number S of species expected to be observed in a geographical zone with area A. The most common SAR is called the Arrhenius relationship (Arrhenius, 1921; Preston, 1948) and posits that S is proportional to a power of A:

$$S = cA^z,$$

where c, z are constants. This relationship is not supposed to hold at all spatial scales. Rather, it is generally believed that SARs are triphasic (Hubbell, 2001; Rosindell and Cornell, 2007) and follow an inverted S-shape when plotted on a log-log scale, displaying a linear intermediate phase with slope z. Many other models of SAR exist, see for example Tjørve (2003) for a review.

• **The Beta-Diversity** is one of Whittaker's indices of biodiversity (Whittaker, 1960), representing the relation between local and global diversity. It is sometimes quantified by the probability that two individuals sampled randomly at a distance r belong to the same species (Chave and Leigh Jr, 2002; Zillio et al., 2005; O'Dwyer and Green, 2010).

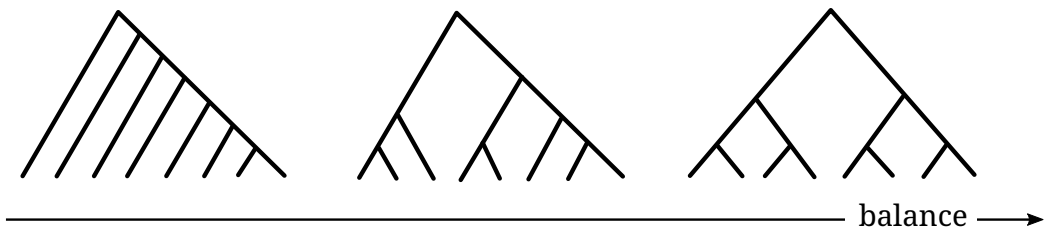
• **The Range Size Distribution** gives the proportion of observed species with a given range size. Speciation rate is thought to be one of the major determinants of range size. However, the relation between range size and speciation rate shows conflicting evidence (Gaston, 1996).

bioRxiv preprint doi: https://doi.org/10.1101/2023.11.09.564915; this version posted November 10, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Box 4: **The shape of phylogenetic trees.**

In order to compare phylogenies, one can use a distance function on the space of phylogenetic trees, such as the Robinson–Foulds distance (Robinson and Foulds, 1981) or one of the many available alternatives (see e.g. Kuhner and Yamato 2015). However, in order to uncover general patterns – or simply to focus on specific aspects of phylogenies – it is often convenient to use summary statistics, also referred to as *shape statistics* in that context.

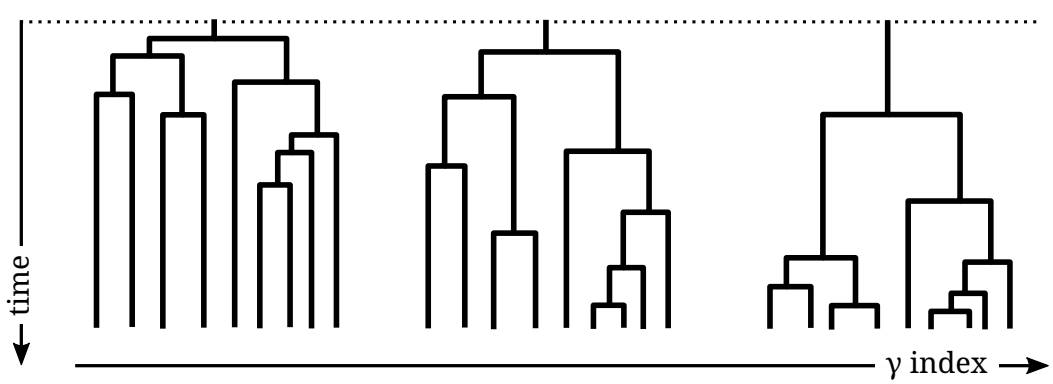
• **Balance indices** are one of the most important classes of such shape statistics. Their goal is to quantify the intuitive idea that some trees are more "balanced" or "have more symmetries" than others. For instance, in the example below, the complete binary tree on the right conforms more to our idea of what it means to be "balanced" than the caterpillar tree on the left.



The best known balance indices are perhaps the Colless index (Colless, 1982) and the Sackin index (Shao and Sokal, 1990); but there are about 20 balance indices whose mathematical properties and usefulness in applications have been studied extensively over the past decades, see Fischer et al. (2021 preprint) for a recent survey. In addition to these indices, an alternative approach to quantify the balance of a phylogeny is to fit a model whose parameter is expected to correlate with some notion of balance. Two prominent examples of this are the parameter α of Ford's model (Ford, 2006) and the parameter β of Aldous's β-splitting model (Aldous, 1996).

One of the uses of balance indices is to test whether new species appear at a higher rate in some parts of a phylogeny than in others. Indeed, models in which all lineages branch at the same rate, such as birth-death processes, yield very balanced phylogenies. Remarkably, most phylogenies encountered in practice turn out to have a comparable degree of balance: significantly lower than that of birth-death trees ($\beta = 0$) but significantly higher than that of uniform ("proportional to distinguishable arrangements") trees ($\beta = -3/2$), often close to a β-splitting model with $\beta = -1$ (Blum and François, 2006). This is an example of a macroevolutionary pattern that has yet to be fully explained.

• **The γ index and LTT plots** belong to another important class of shape statistics that aim to measure "how early *vs* late" most speciation events occurred (or, more generally when branch lengths do not correspond to physical time, how "close to the root" the nodes of the tree are). For instance, in the time-embedded phylogenies below, speciations tend to occur earlier in the phylogeny on the left than in the phylogeny on the right.



bioRxiv preprint doi: https://doi.org/10.1101/2023.11.09.564915; this version posted November 10, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

The best known statistic to capture this idea is the γ index introduced by Pybus and Harvey (2000); but, to get a more complete picture, lineage-through-time (LTT) plots are frequently used (Harvey et al., 1994). Unlike balance indices, which depend only on the topology of the tree, the γ index and LTT plots use the branch lengths. Moreover, whereas balance indices quantify how uniformly speciation events are distributed "horizontally" in the tree, the γ index and LTT plots provide information about their "vertical" distribution in the tree; they can therefore be used to test, e.g, whether speciation rates are time-homogeneous. In particular, a negative value of γ is often observed in practice and can be interpreted as a diversification slowdown (Moen and Morlon, 2014). This is yet another example of a macroevolutionary pattern that is not fully understood.

• **Phylogenetic diversity** and related statistics aim to infer biodiversity from phylogenies. Again, there are many options to choose from (see e.g. Schweiger et al., 2008), but the most commonly used statistic is the PD index introduced by Faith (1992). It is simply the total branch length of the tree – with the idea that diversity accumulates linearly with time in each lineage. Besides its use in phylogenetics, the PD index has been used in conservation, as it can help us understand, e.g, how much biodiversity is lost as a result of species extinction (Mooers et al., 2012; Lambert and Steel, 2013).

2 A review of archetypal models of neutral speciation

2.1 Clonal evolution models

The simplest archetypal model of speciation is the *point mutation model* of Hubbell's *unified neutral theory of biodiversity and biogeography* (UNTB, Box 5). Hubbell uses the notion of *ecological drift* (neutral competition between individuals of different species) to describe the biodiversity of a community of species at the same trophic level using an individual-based model that exactly mirrors fixed-population-size neutral models of population genetics, i.e. Moran or Wright–Fisher models with mutation, under the infinite-allele assumption (see Section 1.3 and Figure 2). This model reconciles Fisher's logseries (Fisher et al., 1943b) and Preston's lognormal distributions (Preston, 1948) for species abundances (see Box 3), as it predicts a SAD that approaches one or the other distribution as the three parameters of the model vary.

Notwithstanding some skepticism about the neutrality assumption (see references in Etienne et al. 2007), this theory has been criticized for the lack of detail in the way it models speciation – notably the fact that in UNTB, speciation is instantaneous and occurs at a rate exactly linear in species abundance. Etienne and his co-authors have proposed more flexible versions of Hubbell's initial model: Etienne et al. (2007) proposed a modification of the theory where the speciation rate is constant across species, as in the birth-death models introduced in Section 1.1, and Etienne and Haegeman (2011) introduced a model of speciation by *random fission* – a first attempt at modeling allopatric speciation, see Box 5. Interestingly, both models failed at improving the fit to SAD data previously provided by the point mutation model. Assumptions on the mode of speciation thus have an important impact on species abundance distributions – see Kopp (2010) for a more complete review on this subject.

With the increasing amount of phylogenetic trees being reconstructed through molecular methods, Jabot and Chave (2009) proposed to take phylogenies into account when fitting data to Hubbell's neutral model. Because the so-called fundamental biodiversity number θ (the rescaled mutation rate, see Box 5) impacts the balance of phylogenies (see Box 4), their method improved the estimation of the parameters of UNTB compared to existing methods. However, it has been argued that this model makes unrealistic predictions about species lifetimes, speciation rates and

bioRxiv preprint doi: https://doi.org/10.1101/2023.11.09.564915; this version posted November 10, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

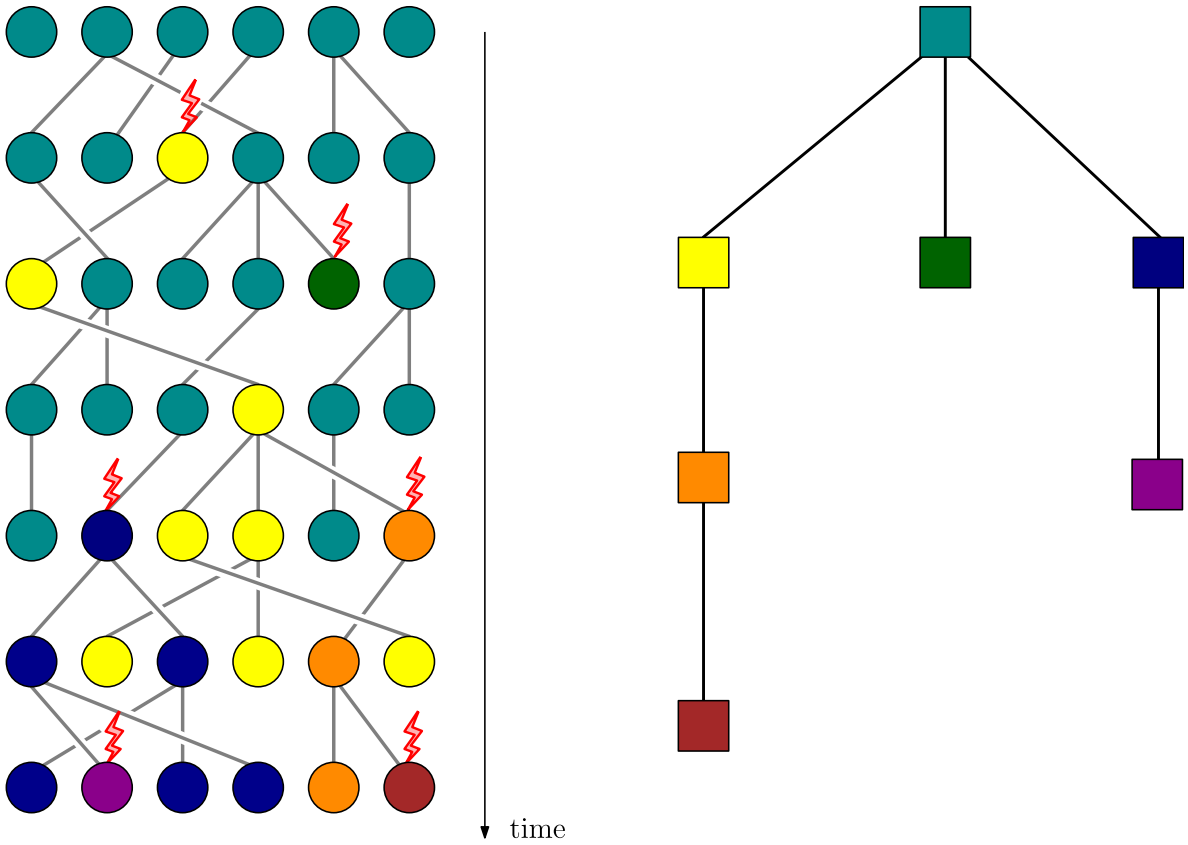


Figure 2: An individual-based genealogical model with mutations under the infinite-allele model. Left: each circle represents an individual, and each color an allelic type. A red lightning indicates a mutation, yielding a new type. Right: the allelic tree corresponding to the process on the left.

number of rare species by Rosindell et al. (2010), who introduced a model where speciation is not instantaneous. This model, known as the *protracted model of speciation*, is a multistage model where species are formed of a dynamic swarm of populations that are able to evolve into so-called "good" species after surviving long enough (see Box 6). This model yields realistic predictions for quantities related to species lifetimes, as well as for the SAD. Its predictions on phylogenies have been studied by Etienne and Rosindell (2012); Etienne et al. (2014) and Lambert et al. (2015), who proposed it as an explanation for the diversification slowdown near present-time observed in real phylogenies. In particular, Etienne et al. (2014) showed that the protracted birth-death diversification model correctly estimates speciation times from phylogenies.

An advantage of the previous models is that the genealogy of the individuals and the clustering of the population into species (see Section 1.3) are constructed jointly. However, the resulting species partitions are not monophyletic in general (Manceau et al., 2015; Manceau and Lambert, 2019), i.e. a species cannot always be defined as the group consisting of all descendants from a single ancestor. In order to circumvent this issue, Manceau et al. (2015) proposed another definition of species in this context, yielding a coarser clustering into species, with a method leading to what they called a model of *speciation by genetic differentiation*. Given a set of point mutations distributed across a genealogy, species are defined as the smallest monophyletic groups of individuals such that any pair of individuals carrying the same genotype are always in the same species. With this criterion, speciation takes time, as mutations arising in a large species will generally not create a new species instantaneously – a property that makes a bridge with protracted speciation models.

bioRxiv preprint doi: https://doi.org/10.1101/2023.11.09.564915; this version posted November 10, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

The authors found that this model yields realistic phylogenies – i.e. with β and γ statistics (see Box 4 and references therein) close to those observed in real data, provided that communities are assumed to be expanding – that is, when the constant-population-size (also called zero-sum) assumption of Hubbell's UNTB is relaxed.

Box 5: **The Unified Neutral Theory of Biodiversity.**

The *Unified Neutral Theory of Biodiversity* (UNTB) is a conceptual framework introduced by Hubbell (2001) that regroups several closely related models. In its simplest form, it describes the diversity and abundances of species on two scales: in a *local community* of size J; and in a much larger *metacommunity* of size J_M.

The dynamics of the metacommunity are governed by two phenomena:

- **Ecological drift.** Each individual dies at constant rate and is then replaced by the offspring of another individual sampled uniformly at random.
- **Speciation by point mutation.** Each individual mutates and forms a new species at rate μ.

The equilibrium distribution for the metacommunity displays a logseries SAD, where the *fundamental biodiversity number* $\theta = 2\mu J_M$ takes the role of Fisher's α parameter. This equilibrium state is known as Ewens's sampling formula in population genetics.

Assuming that the mutation rate is small – i.e. $\mu = \theta/(2J_M) \ll 1$ – and that the local community is much smaller than the metacommunity ($J \ll J_M$), it becomes possible to neglect speciation in the local community. Its diversity then comes through immigration from the metacommunity, which acts as an external buffer. More specifically, the dynamics of the local community are governed by the following events:

- **Ecological drift and immigration.** Each individual dies at constant rate and is then replaced by the offspring of another individual. With probability m, this individual is sampled uniformly in the metacommunity; with probability $1 - m$, it is sampled uniformly at random in the local community.

The equilibrium species abundance distribution in the local community is called the zero-sum multinomial distribution. The lower the immigration parameter m, the more isolated the local community and the fewer rare species (i.e. species with few individuals) at equilibrium, which results in a lognormal-like SAD.

In this model, when new species appear they only consist of a single individual. In Hubbell (2001, 2003) and Hubbell and Lake (2003), Hubbell also proposed alternative modes of speciation:

- **Speciation by random fission.** When an individual undergoes a speciation event, its whole species, of size N, is divided into two new species: one of size K, where K is a uniform random variable between 1 and N; and the other of size $N - K$. This mode could represent a random geographic barrier that splits the species in two.
- **Peripatric speciation.** An intermediate mode between point mutation and random fission, where species start with a small number of individuals.

bioRxiv preprint doi: https://doi.org/10.1101/2023.11.09.564915; this version posted November 10, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Box 6: **The protracted model of speciation.**

Rosindell et al. (2010) criticized previous models of neutral speciation for failing to take into account the fact that speciation takes time. Their model of *protracted speciation* addresses this by considering a parameter τ that corresponds to the time it takes for a so-called *incipient species* – i.e. a subpopulation of a species that has started to differentiate – to turn into a species in its own right (a so-called *good species*). Behind this transition period are hidden complex biological processes that the authors do not model explicitly (see Section 3.1).

The basic model is an extension of Hubbell's UNTB model described in Box 5. The dynamics of the local community are unchanged, but the dynamics of the metacommunity are replaced by the following process:

- **Ecological drift.** Each individual dies at constant rate 1 and is then replaced by the offspring of another individual selected uniformly at random.
- **Differentiation by point mutation.** Each individual mutates at rate μ: it is replaced by an individual forming a new incipient species.
- **Protracted speciation.** After a transition time τ, an incipient species becomes a good species.

In this model, μ is to be interpreted as a "speciation initiation" rate, and the true speciation rate is $\mu_{\text{eff}} = \mu/(1+\tau)$. In the metacommunity, the model predicts an expected number of species with abundance n of

$$\mathbb{E}[S_n] = \frac{\theta}{n}\left(\left(1-\frac{\mu}{1+\tau\mu}\right)^n - \left(1-\frac{1}{1+\tau}\right)^n\right),$$

where $\theta = \mu J_M/(1+\mu)$ – this yields a logseries-like SAD for small values of τ and a lognormal-like SAD for large values of τ.

Protracted birth-death models. In the setting of growing populations, another approach to protracted speciation – proposed in Etienne and Rosindell (2012) and further studied in Etienne et al. (2014); Lambert et al. (2015) – makes the time to speciation a random time: a species gives birth to an incipient species at rate λ_1, which turns into a good species at rate λ_2. The authors show that these models yield satisfying predictions about the shape of phylogenetic trees, in particular they can explain the diversification slowdown near the present (Etienne and Rosindell, 2012), and they accurately predict speciation times (Etienne et al., 2014).

2.2 Models of genetic isolation

In this second category of models, we explicitly consider whether two individuals or populations are reproductively compatible. Doing so allows us to define species directly from interfertility relationships, e.g. as the connected components of the graph whose vertices are populations and edges represent interfertility (see Box 1). Bienvenu et al. (2019) considered such a representation explicitly: a random graph in which edges vanish after some random time, mimicking the build-up of reproductive isolation; and evolving under extinction-recolonization dynamics, where each newborn population inherits the edges of its mother. They make accurate predictions on the degree of a vertex and the number of species, but further work is needed to provide joint predictions for phylogenies and species abundances.

Most other models build on clonal evolution models and on additional assumptions on the genetic basis of reproductive isolation. In some, but not all, in addition to mutations differentiating gene pools, gene flow (sexual reproduction, migration between demes) can tend to homogenize them. This gene flow can be maintained as long as the dissimilarity between individuals

bioRxiv preprint doi: https://doi.org/10.1101/2023.11.09.564915; this version posted November 10, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

or populations remains low enough. In the most complex models, increasing dissimilarity due to mutations is assumed to in turn reduce gene flow, potentially leading to reproductive isolation. One could also consider cases where dissimilarity enhances gene flow (e.g., disassortative mating, self-incompatibility) but these models are not studied here.

We distinguish between two subcategories of models depending on assumptions on the genetic basis underlying the reduction of gene flow (see Box 2).

2.2.1 Reproductive isolation as a by-product of genetic distance

The main example of an archetypal model of distance-based speciation with gene flow is the *Higgs–Derrida model* (HDM) of speciation (see Box 7). Higgs and Derrida (1992) considered a panmictic population of fixed size, with a large number of loci undergoing mutation and recombination, and proved that speciation could occur without selection nor geographic structure. Here, reproductive isolation occurs when the genetic distance exceeds a threshold, see Section 1.3 and Boxes 1, 2 and 7.

It is natural to embed such a model (Higgs and Derrida, 1991; Nei et al., 1983) in a metapopulation composed of demes connected by migration. For example, Manzo and Peliti (1994) have extended the HDM to a model where individuals live on islands connected by rare migration and showed that speciation is more likely in their model than in the original sympatric HDM. Using individual-based simulations of a model similar to the one considered by Manzo and Peliti (1994), Gavrilets et al. (1998, 2000b) proved that gene flow does not impede speciation, even in the absence of any mechanism favoring divergence like differential adaptation. Gavrilets (1999) analytically studied the evolution of the mean genetic distance within and between subpopulations in a stepping-stone model, under multiple geographic scenarios (isolated populations, populations linked by migration, peripheral population).

The assumption of rare mutations and rare migrations is commonly used to overcome difficulties caused by the potentially large number of coexisting genotypes, as it makes it possible to separate timescales and to neglect variations within populations. The resulting drastic reduction of dimension allows for exact analytic approaches, leading to fruitful predictions. For example, Gavrilets (2000) proposed a deterministic approximation of the dynamics of the genetic distance between the mainland and a peripheral population subject to immigration, assuming that the population is at all times monomorphic with respect to the loci controlling incompatibilities. This approach makes it possible to estimate the speciation time, defined as the time when the genetic distance reaches a predefined threshold, depending on model parameters (migration and mutation rates, threshold value).

A similar methodology has been adopted by Yamaguchi and Iwasa (2013), who introduced a stochastic extension of Gavrilets's deterministic model that accounts for the randomness of migrations and mutations. Notably, even in cases where the deterministic approximation suggests that the genetic distance will stabilize below the speciation threshold – so that migration is strong enough to prevent speciation – the occurrence of a rare sequence of migration events could potentially drive the population towards genetic incompatibility.

Most metapopulation models focus exclusively on a small number (2 or 3) of islands. Miró Pina and Schertzer (2019) introduced a generalization of the model of Yamaguchi and Iwasa to a general metapopulation model, to understand how more intricate geographical constraints might impact the aforementioned predictions. One conclusion of this approach is that the steady-state structure of the metapopulation into species is heavily influenced by the quantity of potential migration pathways connecting any two islands, and that a greater number of clusters or geographical bottlenecks are more likely to facilitate speciation events.

bioRxiv preprint doi: https://doi.org/10.1101/2023.11.09.564915; this version posted November 10, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Box 7: **The Higgs–Derrida model of speciation.**

Higgs and Derrida (1991, 1992) introduced a model where a single population of size M is studied, subject to sexual reproduction and mutation. Each individual of this population bears a genome g of size L, which consists of a family of "spins" $(A_1^g, A_2^g, \ldots, A_L^g)$, with $A_i^g \in \{-1, 1\}$ for all i. Generations are non-overlapping and the population size is constant, similarly to the Wright-Fisher model.

This model takes into account the genetic distance between each pair of genomes g and g' using the notion of *overlap* $q^{g,g'} = 1 - 2d(g, g')$, where $d(g, g')$ is the distance between g and g' defined as:

$$d(g, g') = \frac{1}{2L} \sum_{i=1}^{L} |A_i^{g'} - A_i^g|,$$

which equals $\frac{1}{2}$ (in mean, and exactly when L is brought to infinity) when the two genomes are independent, and 0 when the two genomes are equal.
The genomes of a new generation of individuals are generated using a few key ingredients:

- **Sexual reproduction:** At each generation, each individual, independently, chooses two parents whose genomes are at a distance lower than a fixed threshold $d_{\min} = \frac{1-q_{\min}}{2}$.
- **Free recombination and mutation:** At each locus independently, the individual inherits the allele of one of its two parents with probability $\frac{1}{2}$. This allele then mutates with probability μ.

Species are then defined by a threshold clustering algorithm with threshold $d_{\min}$ (see Section 1.3).

Higgs and Derrida (1992) numerically simulated the evolution of the overlaps when L is large, starting from a homogeneous population. In particular, they were able to make estimates on:

- **Relation between species abundance, mutation rate, and speciation:** If $d_{\min} < d_0(m) := \frac{2\mu m}{1+4\mu m}$, then any species of size m is going to split into two species. Therefore the condition to get at least one speciation event is $d_{\min} < d_0(M)$.
- **Inter/intra-specific diversity:** For each pair of species A and B of respective sizes m_A, m_B, one can define $Q^{AB} = \frac{1}{m_A m_B} \sum_{g=1}^{m_A} \sum_{g'=1}^{m_B} q^{g,g'}$ the mean pairwise overlap between A and B. Taking $A = B$ gives an idea of the intraspecific diversity of A, whereas $A \neq B$ gives an idea of the interspecific diversity between the two species (diversity decreases with overlap). Another indicator of intraspecific diversity is the function σ_A which records the mean number of pairs of individuals which are not able to interbreed inside species A. Simulations indicate that σ_A stays close to 0 most of the time, except during short speciation events. This seems to validate the choice made for the definition of species in this model.

2.2.2 Reproductive isolation due to genetic incompatibilities

In the classical representation of the two-locus Dobzhansky–Muller incompatibilities, two allopatric populations start out with the same genetic composition $AABB$ (see Box 2). Most models focusing on this setting study the waiting time before the appearance of each mutant a and b in its background and it fixation (to $aaBB$ and $AAbb$, respectively), or the conditions under which total isolation is eventually completed in spite of gene flow (Bank et al., 2012; Blanckaert and Hermisson, 2018; Gavrilets, 1997; Gavrilets and Gravner, 1997).

Most generalizations of this model to L partially interacting loci also focus on two populations in allopatry and the time to speciation in the absence of gene flow (Orr and Turelli, 2001; Kondrashov, 2003). Other studies look at the genomic signature of the process (Fierst and Hansen, 2010; Flaxman et al., 2014), the influence of different incompatibility scenarios, and the likelihood of the snowball effect (Cutter, 2012; Livingstone et al., 2012; Gourbiere and Mallet, 2010; Fraïsse et al.,

bioRxiv preprint doi: https://doi.org/10.1101/2023.11.09.564915; this version posted November 10, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

2014).

In all these models, the only steady outcome is isolation between all demes. It is not known what can happen in a setting where the incompatibilities themselves are dynamic, so that new species can constantly arise. Marin et al. (2020) made a first attempt at modeling this, by assuming that introgression of a gene into a population can occur with a probability that increases with the number of co-adapted alleles carried by all other loci in the receiver genome. The main assumption is that each novel allele arising by mutation is co-adapted with all alleles of the background where it arose (or is purged by purifying selection) – but with no other allele. This embodies the idea that a novel allele needs to be "tested" against the genome harboring it. The model enforcing this assumption, called *gene-based diversification model* (GBD), makes some specific predictions on the joint evolution of gene and species lineages, acknowledging that gene trees can greatly differ from each other (and, therefore, from the species tree).

2.3 Models of isolation by distance

A key element missing from most of the previous models is the explicit geographical context of speciation. Spatially explicit models are of particular interest for predicting and understanding empirical observations such as species-area relationships and range size distributions (see Box 3).

A first way to construct a spatially explicit, neutral model of speciation is to extend an existing non-spatial model, such as those introduced by the UNTB (Box 5). This can be done for example by introducing a dispersal kernel, as in Durrett and Levin (1992), who used a contact process, i.e. a stochastic nearest-neighbour epidemic process on the lattice , which is a natural choice for modeling dispersal. In their models, new species are introduced by immigration from outside the system or by speciation events, the latter being modeled by point mutations. Their objective was to make sense of Arrhenius SAR – i.e. of the fact that the number of species is proportional to a power of the area (MacArthur and Wilson, 1967, see also Box 3). The model predicts that the exponent in the SAR depends on the rate of introduction of new species, which can explain the variability observed across different taxa and locations. Subsequently, Chave and Leigh Jr (2002); Zillio et al. (2005) used models of neutral biodiversity with limited dispersal close to the one of Durrett and Levin (1992) to predict species beta-diversity (defined there as the ratio between regional and local species diversity) in tropical forests. Rosindell and Cornell (2007) also used a spatially explicit version of Hubbell's neutral model where they varied the dispersal range and the dispersal kernel. They found that the model always predicted the same power law for the SAR, up to rescaling of the two axes, and showed that it exhibits the triphasic behavior that is empirically observed in data. Using fat-tailed dispersal kernels in Rosindell and Cornell (2009), they improved the fit for the SAR, finding realistic values for the exponent of the Arrhenius relationship (Box 5). In O'Dwyer and Green (2010), a first analytical prediction of the triphasic behavior of the SAR in a neutral setting was found, using quantum field theory, along with a prediction of the beta-diversity close to the one of Chave and Leigh Jr (2002). Moreover, they were able to link the exponent z in the Arrhenius SAR to parameters of their formula for beta-diversity.

An alternative way to spatialize speciation is to start with a model of genetic isolation, as in the previous section. Desjardins-Proulx and Gravel (2012a) extended the model of Economo and Keitt (2008) based on the interactions within a set of spatially-organized local populations under ecological drift (see Box 5), and where reproductive isolation is modeled by genetic incompatibility (see Box 2). Their neutral model could not match empirical data of species diversity with realistic mutation rates. Using random geometric networks in a "pseudo-selection" model, Desjardins-Proulx and Gravel (2012b) found that species richness was higher in more connected communities, while speciation was facilitated in more isolated communities. This result is similar to the prediction

bioRxiv preprint doi: https://doi.org/10.1101/2023.11.09.564915; this version posted November 10, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

made by Miró Pina and Schertzer (2019).

Other models are extensions of the HDM (Box 7) in which individuals also have a spatial location and reproduce with their neighbors. de Aguiar et al. (2009) and de Aguiar (2017) considered a spatial version of the HDM to get what they named a "topopatric" model of speciation. They also found that SAR and SAD match observations made for different taxa and regions. In their models, individuals can only mate if their spatial distance and genetic distance are lower than fixed values. This is something that Martins et al. (2013) also did, in the case of ring species. Gavrilets et al. (2000a) studied a 1D array of demes undergoing extinction and recolonization, where speciation is modeled by clonal evolution. Focusing on the movement of species boundaries, they obtained analytical predictions for the average number of species and the average species range, showing that the former scales like $\sqrt{\delta/\mu}$, δ being the extinction-colonization rate and μ the mutation rate per deme. Using numerical simulations, they obtained species range distributions that match those observed in empirical data. In these models, reproductive isolation arises when dissimilarity exceeds a fixed threshold. In contrast, Hoelzer et al. (2008) used cellular automata to model explicitly the genomes of individuals on a 2D grid, in a sexually reproducing population with limited dispersal, where the fitness of an offspring is a decreasing function of the dissimilarity of the genomes of its parents. Pigot et al. (2010) modeled how species boundaries move in a more coarse-grained way than Gavrilets et al. (2000a), overlooking the microscopic dynamics of individuals inside species. They considered two modes of speciation: vicariance – when a barrier intersects the range of an extant species – and peripatry – when the individuals in the edge of the range move to a new location. Their predictions on the shape of phylogenetic trees – which are imbalanced and show a slowdown in the diversification rates – match bird phylogenies; but their predictions on the SAR do not match empirical observations.

3 New models and perspectives

In this section, we outline several new archetypal models of speciation, hopefully tractable from a mathematical point of view. Each of these models aims to address a specific question that is only partially addressed by existing models.

The first group of models, which can be seen as refinements of existing protracted speciation models, aim to model **speciation collapse due to secondary contact**. They open the black box of "protracted speciation" in a parsimonious way: although gene flow is not modeled explicitly, they allow for incomplete reproductive isolation between populations that have not fully evolved into distinct species yet, and could therefore merge back into one species.

The second group of models aims to address a shortcoming of the threshold models of reproductive isolation reviewed in Section 2.2. Indeed, these models – in which individuals/populations lose all ability to interbreed once their genetic distance exceeds a certain value – do not incorporate the fact that **dissimilarity feeds back on homogenization**: either two populations are similar and can interbreed; or they are dissimilar and cannot interbreed. The model that we introduce to tackle this issue instead uses a continuous relationship between gene flow and dissimilarity.

The third class of models that we introduce are models that aim to make predictions on **spatial patterns of speciation** – in particular, on the relationship between species range and speciation rates. We consider two such models: a purely neutral one (the *freezing voter model*) and a purely adaptationist one (the *Red Queen model*). Despite the simplicity of their formulation, these models, and especially the freezing voter model, are challenging to study mathematically.

bioRxiv preprint doi: https://doi.org/10.1101/2023.11.09.564915; this version posted November 10, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

3.1 How does speciation collapse slow down diversification?

One of the innovations of the protracted speciation model of Rosindell et al. (2010) has been to recognize and incorporate the fact that speciation is not instantaneous. However, in this model, the process by which so-called *incipient* species turn into *good* species remains a black box, and is reduced to its duration τ, which is set as an exogenous parameter of the model. In a more realistic context, the transition from incipient to good species is the result of a complex interplay between differentiation and homogenization. This process not only unfolds over time, but also carries the potential for failure – resulting in what is referred to as *speciation collapse*.

To gain a comprehensive understanding of this effect, it becomes essential to enhance the initial protracted speciation model with a thorough description of the competition between differentiation and homogenization. This will make it possible to obtain the transition time τ of Rosindell et al. (2010) as an emergent property, and at the same time to assess the role of speciation collapse in speciation.

A simple model consists in considering a collection of populations where differentiation is driven by point mutations; homogenization happens as a result of reproductively compatible populations merging; and populations become reproductively isolated past a certain threshold. More specifically, one can for instance consider a set of populations carrying types, where:

- Populations can split or become extinct as in a birth-death process;
- Each population mutates at rate μ, taking a new type never seen before (infinite-allele model).
- Each pair of reproductively compatible populations merges at constant rate, where:
  - Two populations are reproductively compatible if their types are at distance less than k in the genealogy of the types (see Section 1.3).
  - When two populations merge, the type of the resulting population is chosen according to some specific rule. For instance, the merged type could be one of the two parent types chosen at random; or it could be a new type, seen as a child of the two parent types – in which case the genealogy of the types is given by a directed acyclic graph instead of a tree – see Figure 3.

Like protracted speciation models, this model describes the dynamics of a swarm of evolving populations. In protracted speciation models, these populations are purely diverging, and this divergence is artificially described as a two-step process (incipient species turning into good species). Here, not only is there a possibility for homogenization, but the divergence is also a more gradual process: there is no need for a notion of incipient species. However, to fall back on the protracted speciation setting, each new type can be interpreted as an incipient species.

The case $k = 1$, where only populations of the same type are reproductively compatible, has been studied in detail in Bienvenu and Duchamps (2022 preprint). Even though, like the protracted speciation model of Rosindell et al. (2010), this specific case lacks the crucial ingredient that is homogenization, it provides a helpful guide to see the kind of mathematical results that one can hope to obtain for such models, as well as the challenges that their study poses. In particular, for $k = 1$ differentiation is always sufficient to ensure speciation – in the sense that the network describing the genealogy of the populations has a tree structure on the large scale; moreover, this tree structure is fully understood (namely, it is the continuum random tree of Aldous, 1991). But whether this is the case for every value of k is an open and challenging mathematical question.

Another very natural way to model the competition between differentiation and homogenization is to assume that populations diverge at a constant rate and that homogenization is therefore

bioRxiv preprint doi: https://doi.org/10.1101/2023.11.09.564915; this version posted November 10, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

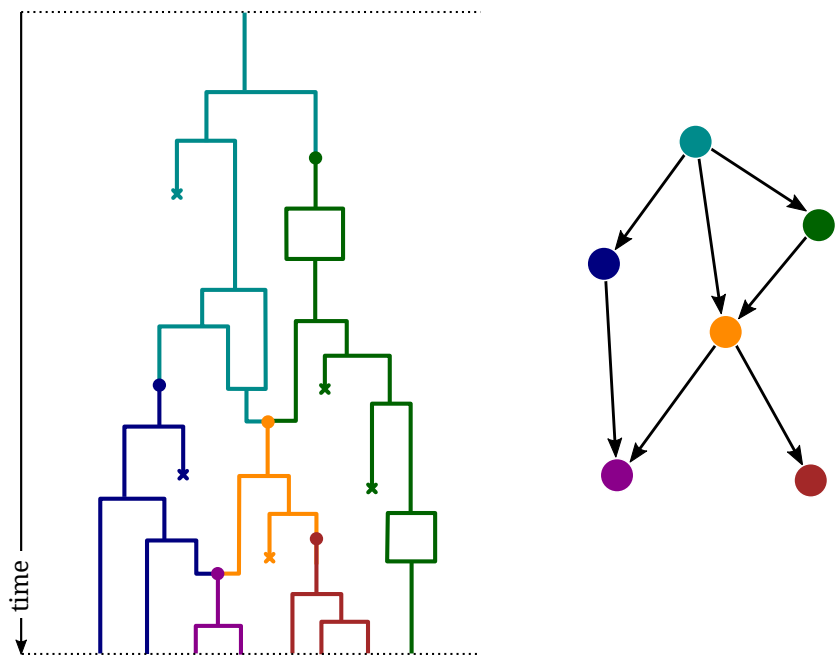


Figure 3: Left, a realization of the process described in the main text. Each vertical line corresponds to the lineage of a population, and each color to a type. Right, the directed acyclic graph encoding the genealogy of the colors. In this example, populations are reproductively compatible if and only if their colors are separated by at most two edges in the genealogy of the colors.

a decreasing function of their divergence time. Formally, one can consider a model where, as previously, populations branch and die at a constant rate; and where two populations i and j merge at rate $f(t_{\mathrm{MRCA}}(i,j))$, where $t_{\mathrm{MRCA}}(i,j)$ is the time to a most common ancestor of i and j, and f is a decreasing function parametrizing the model.

As for the previous model, we say that there is speciation when the network describing the genealogy of the populations has a tree structure on the large scale (i.e. when there are groups of populations that have diverged so much that their descendants coexist without ever merging together). The outcome of the competition between differentiation and homogenization – i.e. whether the homogenization, as parametrized by a given function f, is strong enough to prevent speciation – is highly non-trivial. In the case where f decreases exponentially, i.e. $f(t) = \beta e^{-\alpha t}$ for some positive constants α and β, one can prove that there is speciation for a large range of parameters (in particular, whenever $\beta > \beta_c \approx 1.26$ or $\alpha \geq 2$). Simulations – see Figure 4 – and heuristics suggest that (i) speciation occurs when $f(t) = \beta e^{-\alpha t}$ for all positive values of α and β but (ii) speciation never occurs if f decreases more slowly than exponentially. However, giving a rigorous proof of these statements – and, *a fortiori*, characterizing the shape of the resulting phylogeny – are wide-open questions.

3.2 How does dissimilarity slow down homogenization?

In order to motivate the next model, we recall that in many of the models presented in Section 2.2.1, reproductive isolation was presented as a threshold effect: two individuals are reproductively compatible if their genetic distance is below a predetermined threshold; otherwise they are reproductively incompatible.

In reality, speciation is inherently characterized by a gradual transition, and there is a need to explore the potential outcomes of depicting reproductive isolation as a continuous function of genetic

bioRxiv preprint doi: https://doi.org/10.1101/2023.11.09.564915; this version posted November 10, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

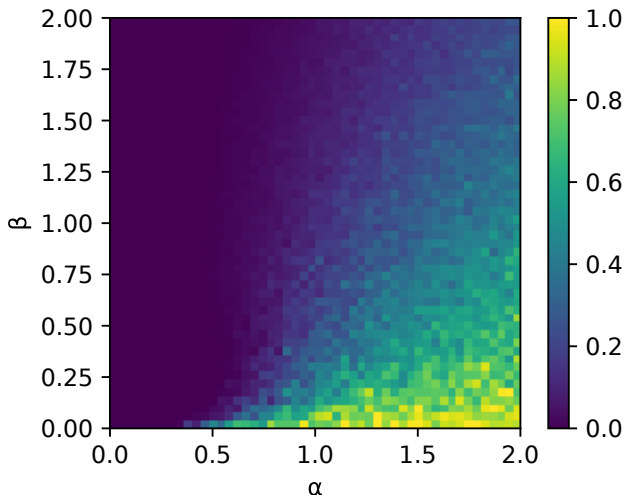


Figure 4: Estimated speciation rates for the model with continuous differentiation described in the main text. Each pixel is estimated from a simulation started from a single population and run for 2500 units of time (or stopped upon reaching 250 populations).

distance. In this program, speciation will not be represented as a sudden, discontinuous process; rather, it will be represented as a gradual fragmentation resulting from the feedback between migration and reproductive isolation. As two populations become more genetically distant, gene flow decreases, consequently leading to an increase in genetic divergence due to the accumulation of new mutations and so on and so forth. Contrary to the threshold model, it is rather unclear whether such a snowball effect will eventually drive two semi-isolated populations into two distinct species.

In order to convince the reader of the potential of such an approach, we revisit the multi-scale model alluded to in Section 2.2.1 in more detail. The population is partitioned into demes connected by migration. In this setting, mutation and migration are rare, so that intra-deme diversity can be ignored. We assume the existence of $L \gg 1$ speciation loci and define the genetic proximity p_{ij} between demes i and j as the fraction of loci where i and j share the same allele. We assume the existence of a *feedback function* h that encodes the degree of reproductive compatibility as a function of genetic proximity. It is natural to assume that h is increasing (i.e. the gene flow between two demes increases with their genetic proximity) and that $h(0) = 0$. Finally,

- In each population and at each locus, new mutations fix independently at rate μ;
- An "effective" migration between i and j occurs at rate $m_{ij}h(p_{ij})$, where m_{ij} is the maximal migration rate from i to j. Upon effective migration, a single allele of the migrant i is fixed in population j.

The feedback function h may be hard to measure in practice, but some of its qualitative features could shed light on potential speciation scenarios.

As an illustration, consider the simple situation of two demes, with genetic proximity $p(t)$ at time t. When $L \gg 1$ and assuming symmetric migration at rate m, the dynamics of p are well approximated by a deterministic equation

$$\frac{dp}{dt} = 2\left(mh(p)(1-p) - \mu p\right) \tag{3}$$

bioRxiv preprint doi: https://doi.org/10.1101/2023.11.09.564915; this version posted November 10, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

To gain some intuition on the derivation, the first term on the right-hand side is the effect of migration on genetic diversity (p increases), whereas the second term is the effect of mutation (p decreases). For a general migration graph, the previous approach can be generalized so that the genetic proximities ($p_{ij}(t)$) are now solution to an explicit, non-linear system of differential equations.

Figure 5 shows the outcome of numerical simulations, where N demes are connected as in a complete graph with symmetric, uniform migration rates. At time $t = 0$, the population starts from a symmetric equilibrium and is exposed to a temporary stress, which is enforced by temporarily lowering the migration rate. We observe that the population's response highly depends on the shape of h. When $h(x) = x$ (Figure 5a), 0 is unstable and the population recovers its initial level of gene flow from any environmental stress. This outcome can be interpreted as a species collapse following secondary contact. In contrast, for $h(x) = x^{3.5}$, which corresponds to a much faster decrease in gene flow with genetic dissimilarity (Figure 5b), an intense stress can drive the population to speciation. In fact, the population transitions from a single species to N distinct species. Beyond these two specific examples, a careful analysis of the limiting deterministic system enables us to connect the resilience of a species (or conversely, the occurrence of speciation) to the behavior of the feedback function h at 0. If $h'(0) > 0$, the system rebounds from any environmental stress (speciation collapse). If $h'(0) = 0$, speciation takes place following sufficiently strong environmental stress.

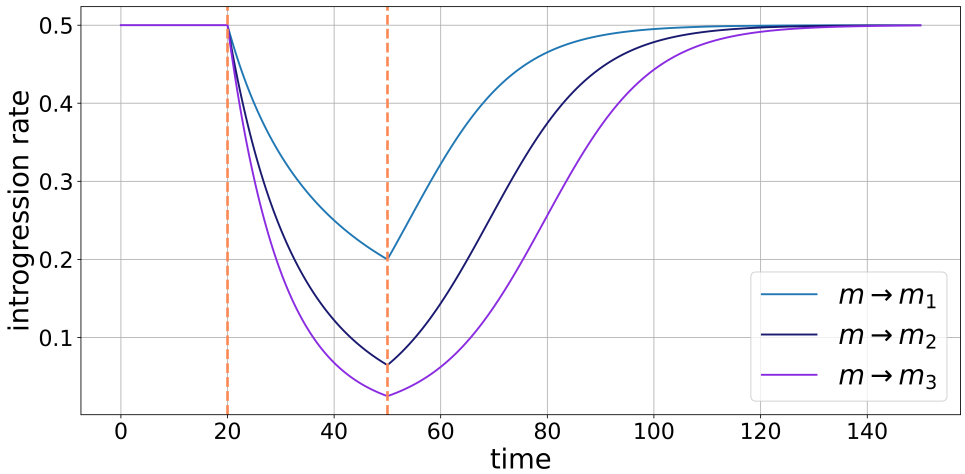
The previous approach can be extended to tackle problems related to the geographical structure of a species complex. For instance: under which conditions do we observe a stable ring species? This can be answered through a generalization of the dynamical approach exposed in Equation (3) where the underlying migration graph is a ring (see Figure 6). It can be shown that the existence of a stable ring species again relates to the behavior of the feedback function h at 0. More specifically, a necessary condition is the existence of a speciation threshold s such that gene flow is interrupted if populations are too dissimilar (i.e. $h(x) = 0$ for $x \in [0, s]$). Under this assumption, it remains an open question to derive the range of parameters (the length of the ring, the threshold s etc.) under which a stable ring configuration exists.

In summary, the previous model suggests that the feedback function h has the potential to yield diverse qualitative predictions on speciation (speciation collapse, ring species). Consequently, it would be intriguing to explore how h arises from a more detailed population description. In this refined model, migration is not portrayed as an instantaneous event but rather as a stochastic process, wherein the genetic material of a single migrant gradually disperses over several generations until the fixation of some of its alleles occurs in a focal population. In turn, this invasion of the migrant's genetic material depends on the specific mechanism of genetic isolation at hand. An interesting question would be to find an approximate relation (if it exists) between a given mechanism of reproductive isolation and the feedback function h. Provided such a program could be achieved, one could then relate speciation patterns to the underlying genetic architectures of Box 2.

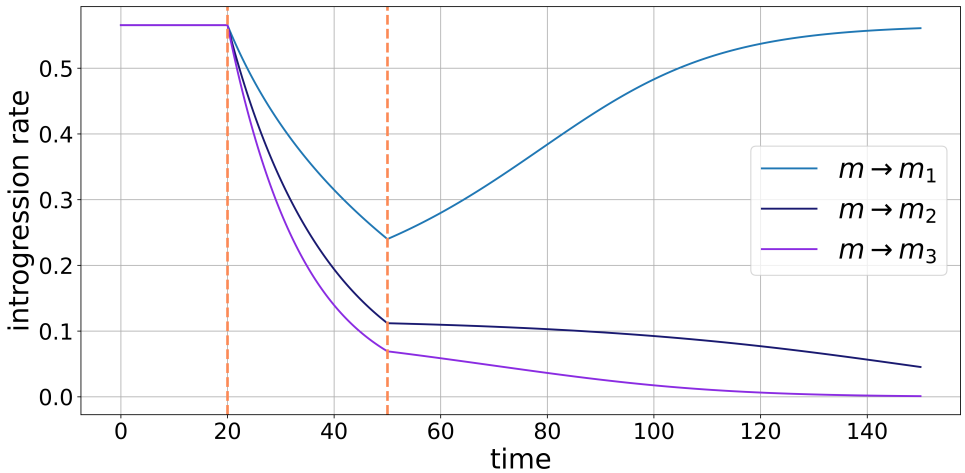
3.3 How does species range control speciation?

Understanding how species range controls the rate of speciation is notoriously difficult. Large species ranges can be thought of as favoring speciation, but a large range can be due to a high dispersal ability, which will in turn increase gene flow and reduce population differentiation, thereby inhibiting speciation. Similarly, spatially fragmented species are seemingly more prone to speciation, but if local populations are too small, they can go extinct before adapting to the local environment. On the empirical side, studies on how speciation is influenced by population size, population structure, species range, evolution of reproductive isolation or population differentia-

bioRxiv preprint doi: https://doi.org/10.1101/2023.11.09.564915; this version posted November 10, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.



(a) $h(x) = x$



(b) $h(x) = x^{3.5}$

Figure 5: We consider a perturbation of a biological system at equilibrium with symmetric migration rates m between populations. The introgression rate (y-axis) is defined as $mh(p(t))$, where $p(t)$ is the genetic proximity at time t. Speciation occurs when the introgression rate becomes very small. The system is exposed to a temporary stress in the time interval delimited by the two vertical orange lines in the form of lowering temporarily m. The original migration rate is reestablished upon releasing the stress. In each plot, the three different scenarios correspond to three different migration changes during the stress period. In the last two cases in (b), speciation occurs. *Parameters:* $N = 6, \mu = 0.01, m = 0.1, m_1 = m/2, m_2 = m/5, m_3 = 0$.

bioRxiv preprint doi: https://doi.org/10.1101/2023.11.09.564915; this version posted November 10, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

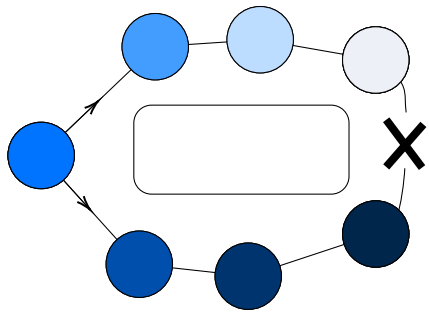


Figure 6: Demes forming a ring around a geographical obstacle. A ring species emerges when two neighboring demes (rightmost subpopulations) cannot interbreed but still exchange genes through intermediary populations.

tion abound: in gastropods (Wagner and Erwin, 1995), mollusks (Jablonski and Roy, 2003), birds (Harvey et al., 2017; Rabosky and Matute, 2013), lizards (Singhal et al., 2018), desert snakes (Alencar and Quental, 2019), freshwater fish (Yamasaki et al., 2020), drosophila (Rabosky and Matute, 2013) or *in silico* (Birand et al., 2012; Maya-Lastra and Eaton, 2021 preprint; Pigot et al., 2010). Overall, empirical findings are not mutually consistent (Harvey et al., 2019; Rabosky, 2016).

A negative relationship between speciation rate and species range has been rediscovered several times (Jablonski and Roy, 2003; Wagner and Erwin, 1995), but again this pattern has ambiguous interpretations – it might be a consequence of speciation dividing ranges and limiting similarity preventing recolonization, rather than an intrinsic property of species with small ranges.

It is therefore important to come up with models that provide predictions on the relationship between range sizes and speciation rates, and with a conceptual framework providing null expectations on how population size, fragmentation and differentiation co-vary and evolve; how they promote or impede speciation; and how they are transmitted to daughter species.

3.3.1 The freezing voter model

This new archetypal model of speciation takes into account very few factors: migration, mutation, and reproductive isolation when the genetic distance between two individuals is greater than a given threshold. It can be seen as a model of co-adaptation restricted to the dynamics at a single locus. When two alleles at this focal locus have accumulated too many differences, they become unable to invade the genetic background of one another.

As in Section 3.2, this model considers N demes forming a graph with edges indicating that migration is possible. Natural choices of graphs include the one-dimensional path (as in the stepping-stone model), for mathematical tractability; and the two-dimensional square grid. But other geometries are possible (for instance, a tree could represent a network of rivers/valleys).

Each deme is occupied by a monomorphic population exchanging migrants with a neighboring deme, having then an opportunity to propagate their type inside the target population. However, the migrant is accepted only if her type is sufficiently close, for example at distance smaller than a given threshold, to the type found in the target population, similarly as in the previous section. Finally, mutations occur as in the infinite-allele model. The distance between two types is equal to the number of mutations needed to go from one type to the other.

We refer to Figure 7 for a schematic illustration of the dynamics of the model, and to Figure 8 for a computer simulation. In this simulation and throughout the rest of the section, we use a threshold equal to 1.

As can be seen in Figure 8, some barriers between genetically distant populations may appear

bioRxiv preprint doi: https://doi.org/10.1101/2023.11.09.564915; this version posted November 10, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

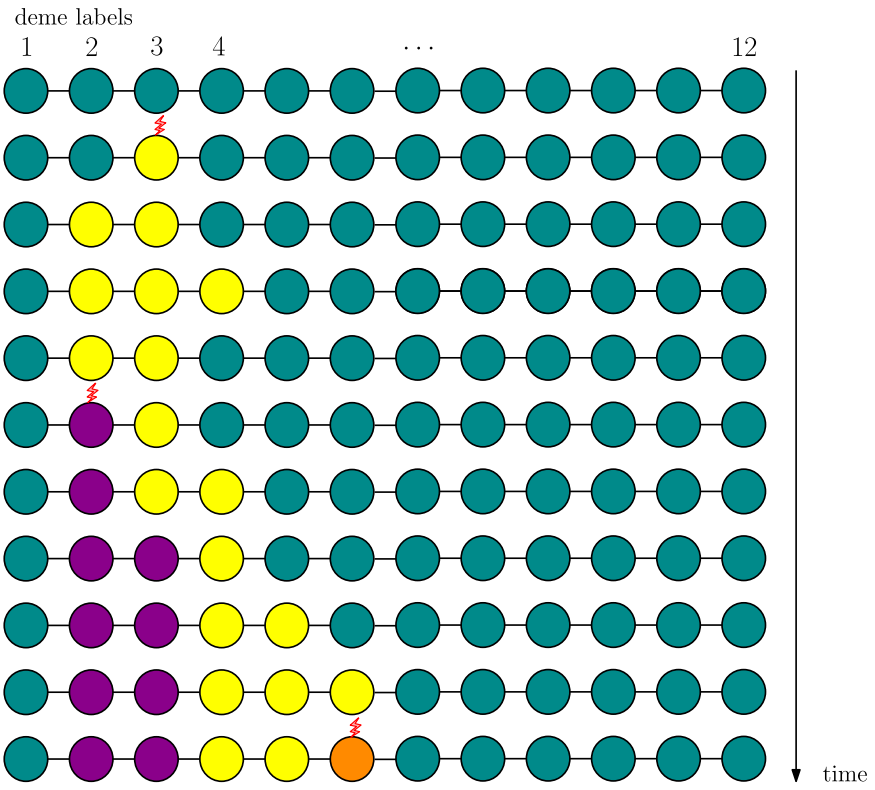


Figure 7: A representation of the freezing voter model, for $N = 12$. Each color represents a type. Each circle is a different deme. Red lightnings indicate mutations.

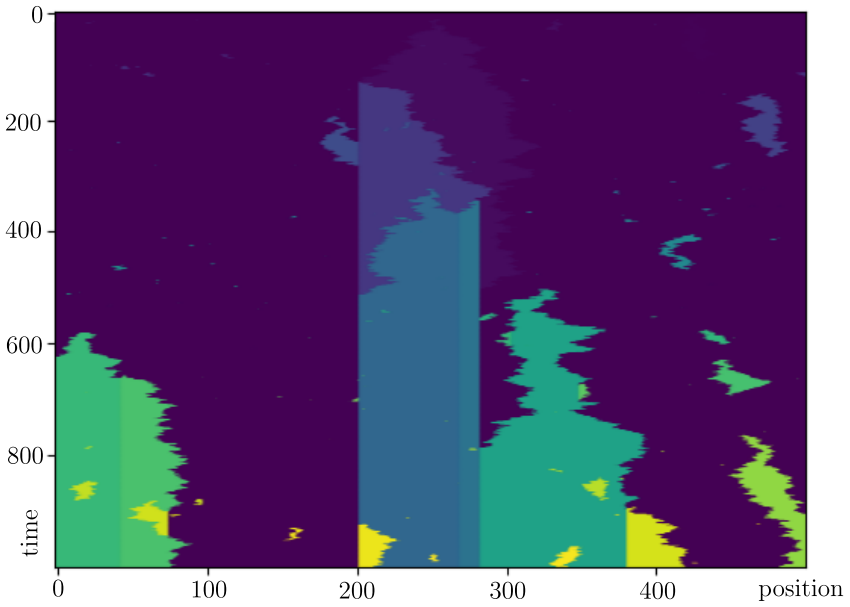


Figure 8: A simulation of the freezing voter model. The horizontal axis represents space. The vertical axis represents time. Each color represents a different genotype. Multiple barriers are formed (vertical lines separating two colors), some of which are temporary (e.g, the barrier between yellow and purple in the bottom left corner) and some of which are permanent.

bioRxiv preprint doi: https://doi.org/10.1101/2023.11.09.564915; this version posted November 10, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

and disappear further in time due to the extinction of one of the two types for the benefit of an "intermediary" type that can interbreed with the other populations. Speciation can be defined here as the formation of a **permanent** reproductive barrier between two demes, that is to say a point in space that will never be crossed by any successful migration due to the threshold condition on genetic distances.

Thinking of types as political opinions, the model can also be interpreted as describing the dynamics of connected individuals sharing their opinions with their neighbors. Individuals sometimes invent new opinions (mutation) and may succeed in convincing their neighbor (migration), unless the two opinions are too different (reproductive barrier). In the case where we start with a fixed number of opinions, no new opinions are created and there are no restrictions on the propagation of opinions, this is the classic *voter model* introduced by Holley and Liggett (1975) – hence the name "freezing voter model" for our model.

This model provides a rigorous framework to study questions such as

- How long does it take for a "barrier" to appear?
- Where do "barriers" appear?
- What is the typical size and shape of a species range? What is the influence of this shape and of this size on speciation rate?

Though mathematically well-posed, these questions are very challenging to study for the freezing voter model. However, they become easier for a variant of this model, which we now turn to.

3.3.2 A simplification: the Red Queen model

A natural variant of the freezing voter model consists in relaxing the assumption of functional equivalence between species. This can be done in a parsimonious way by introducing a simple asymmetry in gene flow by giving an advantage to recent alleles compared to older alleles in the exchange of genes between interbreeding populations. A similar hypothesis was made by O'Dwyer and Chisholm (2014), who made an analogy to the Red Queen hypothesis.

In the continuous-space limit obtained by taking the number of demes to infinity, the range of a population carrying a novel type expands at constant speed until it meets the range of a population at genetic distance greater than some threshold value. More precisely, mutants appear at rate μ at a uniform position on the interval [0,1]; each mutant propagates to its left and right at speed 1 until it meets another mutant whose genotype is at distance more than 1 from its own genotype, at which point a barrier is formed (speciation event). See Figure 9 for an illustration.

In dimension 1 and under the rare-mutation regime ($\mu \to 0$), we are able to find the distribution of the first barrier that forms between two species, that is to say, the location Z of the boundary formed by the first speciation event, and the time T at which it arises. Specifically, for small μ, the speciation time T is well-approximated by an exponential random variable with parameter $\frac{\mu^2}{3}$, and the position Z is well-approximated by a Beta$(2,2)$ variable, i.e. a random variable on $[0,1]$ with density function $f(z) = 6z\,(1-z)$. It is interesting to notice that speciation events are more likely to occur in the center of the range rather than close to its edges, an interesting ecological property known as the "mid-domain effect" (Colwell and Lees, 2000).

Note that Z does not need rescaling to exhibit a limiting distribution, whereas the law of T scales like $\frac{1}{\mu^2}$. This scaling is explained by the fact that, for a speciation to occur, two incompatible genotypes have to meet. Therefore two mutations have to appear in a time-interval of size at most 1, and this happens at rate of order μ^2. For a more precise statement of these results and a proof, see Appendix A, and see Figure 10 for an illustration.

bioRxiv preprint doi: https://doi.org/10.1101/2023.11.09.564915; this version posted November 10, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

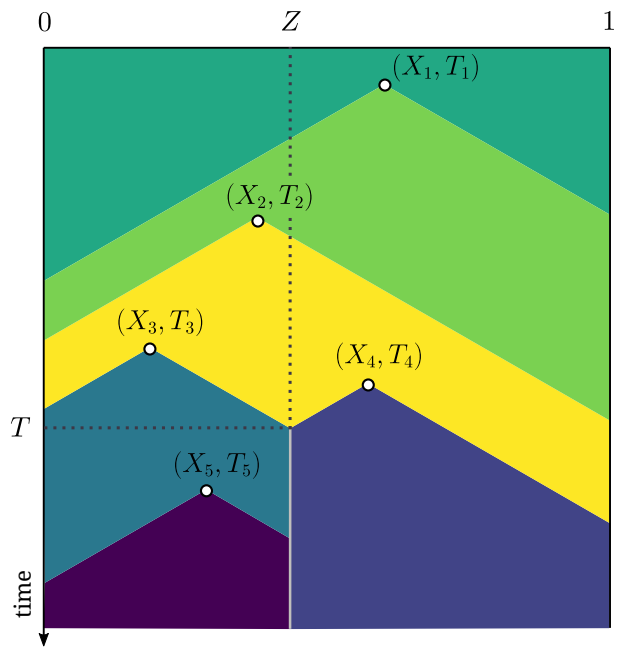


Figure 9: An illustration of the continuous-space limit of the Red Queen model. Time is represented on the vertical axis, and space on the horizontal axis. The (X_i, T_i) are the positions in time and space of the appearances of the mutants. Each color represents a type. Here, a barrier is formed at time T and position Z.

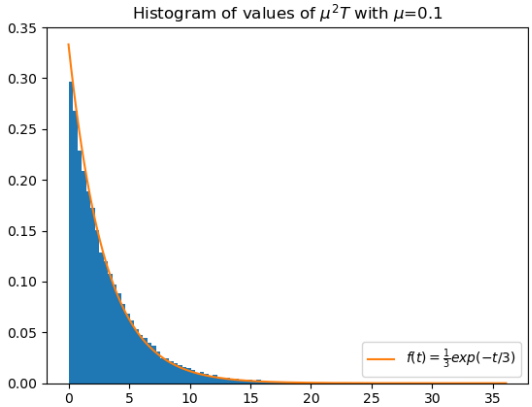
An inductive reasoning allows us to derive the distributions of the subsequent speciation events, using the fact that on two sides of a barrier, the model behaves as two independent models on smaller parts of the space. The shape of the species tree can then be derived from computing these distributions.

Conclusion

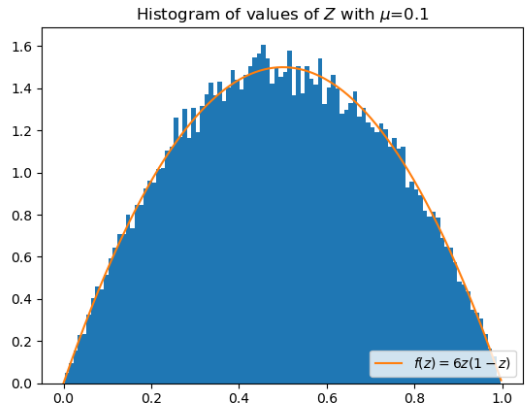
We have set out to demonstrate the advantages of a bottom-up approach to speciation and of what we termed "archetypal" models of speciation – i.e. models that are built from mechanistic principles, have few parameters and focus on specific aspects of the speciation process, disregarding selection forces other than those acting against hybrids. Indeed, macroscopic models and the so-called "lineage-based" approach to diversification have proved very powerful when it comes to portraying macroevolutionary processes; but they are phenomenological in nature and can only shed limited light on the inner causes of macroevolution. By contrast, microscopic models – although typically more challenging to study and harder to relate to real-world data – can, under the right circumstances, yield macroscopic predictions that can be used to assess the validity of our current understanding of evolutionary processes. There is, of course, a trade-off between tractability and realism with this approach, which entails that the goal of an archetypal model can never be to give an accurate description of the evolutionary history of a given taxon, but must instead be restricted to studying specific phenomena in isolation and testing precise hypotheses.

We have reviewed some of the main existing archetypal models of speciation, in order to give an overview of their diversity and of the questions that they can address, but also to point out some of their current limitations. We have identified a set of key mechanisms that are rarely taken into account by these models (the failure of speciation at secondary contact, the feedback of dissimilarity

bioRxiv preprint doi: https://doi.org/10.1101/2023.11.09.564915; this version posted November 10, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.



(a) In blue, the simulated distribution of $\mu^2 T$. In orange, the density of the exponential distribution with parameter $\frac{1}{3}$.



(b) In blue, the simulated distribution of Z. In orange, the density of the Beta$(2, 2)$ distribution.

Figure 10: The variables T and Z are well-approximated by simple distributions: the exponential distribution with parameter $\frac{\mu^2}{3}$ and the Beta$(2, 2)$ distribution whose density is $f(z) = 6z(1 - z)$. In both figures, the sample size is 50000, the mutation rate $\mu = 0.1$.

on homogenization, the emergence in space of reproductive barriers), and we have proposed new models incorporating them. These new models open up promising avenues of research – both for the mathematical challenges that they pose and, more importantly, for the insights into evolutionary processes that their understanding will provide.

A crucial aspect of the micro-macro approach – which we have already mentioned, but which is beyond the scope of this article – is the comparison with real-world data. Indeed, microscopic models are designed based on our current knowledge of evolutionary processes, and in a sense they are the mathematical formulation of our understanding of these processes. Their design and study is an interesting scientific topic in itself; but it is not an end-goal. Instead, the end-goal is to infer their validity from real data, in order to see whether our current understanding of speciation is compatible with reality – and update it accordingly. As of today, there is no consistent relation between the predictions of microscopic models and macroscopic data on the topic of speciation. This reflects the fact that our understanding of the main processes driving speciation and/or the interplay between these processes are not yet fully understood.

Acknowledgments

FB was supported by Dr. Max Rössler, the Walter Haefner Foundation and the ETH Zürich Foundation. AL thanks the Center for Interdisciplinary Research in Biology for funding. We thank Félix Foutel-Rodier, Yannic Wenzel and Philibert Courau for their comments and enlightening discussions.

References

Aldous, D. (1991). The continuum random tree. II. An overview. *Stochastic analysis*, 167:23–70.

bioRxiv preprint doi: https://doi.org/10.1101/2023.11.09.564915; this version posted November 10, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Aldous, D. (1996). Probability distributions on cladograms. In *Random discrete structures*, pages 1–18. Springer.

Aldous, D. and Popovic, L. (2005). A critical branching process model for biodiversity. *Advances in Applied Probability*, 37(4):1094–1115.

Aldous, D. J. (2001). Stochastic models and descriptive statistics for phylogenetic trees, from Yule to today. *Statistical Science*, 16(1):23–34.

Alencar, L. R. and Quental, T. B. (2019). Exploring the drivers of population structure across desert snakes can help to link micro and macroevolution. *Molecular Ecology*, 28(20):4529–4532.

Arrhenius, O. (1921). Species and area. *Journal of Ecology*, 9(1):95–99.

Bank, C., Bürger, R., and Hermisson, J. (2012). The limits to parapatric speciation: Dobzhansky–muller incompatibilities in a continent–island model. *Genetics*, 191(3):845–863.

Bienvenu, F., Débarre, F., and Lambert, A. (2019). The split-and-drift random graph, a null model for speciation. *Stochastic Processes and their Applications*, 129(6):2010–2048.

Bienvenu, F. and Duchamps, J.-J. (2022 preprint). A branching process with coalescence to model random phylogenetic networks. *arXiv:2211.02407*.

Birand, A., Vose, A., and Gavrilets, S. (2012). Patterns of species ranges, speciation, and extinction. *The American Naturalist*, 179(1):1–21.

Blanckaert, A. and Hermisson, J. (2018). The limits to parapatric speciation ii: strengthening a preexisting genetic barrier to gene flow in parapatry. *Genetics*, 209(1):241–254.

Blum, M. G. and François, O. (2006). Which random processes describe the tree of life? A large-scale study of phylogenetic tree imbalance. *Systematic Biology*, 55(4):685–691.

Chave, J. (2004). Neutral theory and community ecology. *Ecology Letters*, 7(3):241–253.

Chave, J. and Leigh Jr, E. G. (2002). A spatially explicit neutral model of β-diversity in tropical forests. *Theoretical Population Biology*, 62:153–168.

Colless, D. H. (1982). Review of "Phylogenetics: The theory and practice of phylogenetic systematics". *Systematic Zoology*, 31:100–104.

Colwell, R. K. and Lees, D. C. (2000). The mid-domain effect: geometric constraints on the geography of species richness. *Trends in ecology & evolution*, 15(2):70–76.

Corbett-Detig, R. B., Zhou, J., Clark, A. G., Hartl, D. L., and Ayroles, J. F. (2013). Genetic incompatibilities are widespread within species. *Nature*, 504(7478):135–137.

Coyne, J. A., Barton, N. H., and Turelli, M. (2000). Is Wright's shifting balance process important in evolution? *Evolution*, 54:307–317.

Coyne, J. A. and Orr, H. A. (1989). Patterns of speciation in Drosophila. *Evolution*, 43(2):362–381.

Cutter, A. D. (2012). The polymorphic prelude to bateson–dobzhansky–muller incompatibilities. *Trends in ecology & evolution*, 27(4):209–218.

bioRxiv preprint doi: https://doi.org/10.1101/2023.11.09.564915; this version posted November 10, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

de Aguiar, M. A. M. (2017). Speciation in the Derrida–Higgs model with finite genomes and spatial populations. *Journal of Physics A: Mathematical and Theoretical*, 50(8):085602.

de Aguiar, M. A. M., Baranger, M., Baptestini, E., Kaufman, L., and Bar-Yam, Y. (2009). Global patterns of speciation and diversity. *Nature*, 460(7253):384–387.

De Queiroz, K. (2007). Species concepts and species delimitation. *Systematic Biology*, 56(6):879–886.

Degnan, J. H. and Rosenberg, N. A. (2006). Discordance of species trees with their most likely gene trees. *PLoS Genetics*, 2(5):e68.

Desjardins-Proulx, P. and Gravel, D. (2012a). How likely is speciation in neutral ecology? *The American Naturalist*, 179(1):137–144.

Desjardins-Proulx, P. and Gravel, S. (2012b). A complex speciation–richness relationship in a simple neutral model. *Ecology and Evolution*, 12(8):1781–1790.

Dobhansky, T. G. (1937). *Genetics and the origin of species*. Columbia University Press.

Durrett, R. and Levin, S. (1992). Spatial models for species-area curves. *Journal of Theoretical Biology*, 179:119–127.

Economo, E. P. and Keitt, T. H. (2008). Species diversity in neutral metacommunities: a network approach. *Ecology letters*, 11(1):52–62.

Etienne, R. S., Apol, M. E. F., Olff, H., and Weissing, F. J. (2007). Modes of speciation and the neutral theory of biodiversity. *Oikos*, 116(2):241–258.

Etienne, R. S. and Haegeman, B. (2011). The neutral theory of biodiversity with random fission speciation. *Theoretical Ecology*, 4(1):87–109.

Etienne, R. S., Morlon, H., and Lambert, A. (2014). Estimating the duration of speciation from phylogenies. *Evolution*, 68(8):2430–2440.

Etienne, R. S. and Rosindell, J. (2012). Prolonging the past counteracts the pull of the present: protracted speciation can explain observed slowdowns in diversification. *Systematic Biology*, 61(2):204–213.

Faith, D. P. (1992). Conservation evaluation and phylogenetic diversity. *Biological conservation*, 61(1):1–10.

Fierst, J. L. and Hansen, T. F. (2010). Genetic architecture and postzygotic reproductive isolation: evolution of bateson–dobzhansky–muller incompatibilities in a polygenic model. *Evolution*, 64(3):675–693.

Fischer, M., Herbst, L., Kersting, S., Kühn, L., and Wicke, K. (2021 preprint). Tree balance indices: a comprehensive survey. *arXiv:2109.12281*.

Fisher, R. A., Corbet, A. S., and Williams, C. B. (1943a). The relation between the number of species and the number of individuals in a random sample of an animal population. *The Journal of Animal Ecology*, pages 42–58.

bioRxiv preprint doi: https://doi.org/10.1101/2023.11.09.564915; this version posted November 10, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Fisher, R. A., Corbet, A. S., and Williams, C. B. (1943b). The relation between the number of species and the number of individuals in a random sample of an animal population. *The Journal of Animal Ecology*, 12(1):42.

Flaxman, S. M., Wacholder, A. C., Feder, J. L., and Nosil, P. (2014). Theoretical models of the influence of genomic architecture on the dynamics of speciation. *Molecular ecology*, 23(16):4074–4088.

Ford, D. J. (2006). *Probabilities on cladograms: introduction to the alpha model*. Stanford University.

Fraïsse, C., Elderfield, J., and Welch, J. (2014). The genetics of speciation: are complex incompatibilities easier to evolve? *Journal of evolutionary biology*, 27(4):688–699.

Gaston, K. J. (1996). Species-range-size distributions: patterns, mechanisms and implications. *Trends in Ecology & Evolution*, 11(5):197–201.

Gavrilets, S. (1997). Evolution and speciation on holey adaptive landscapes. *Trends Ecol Evol.*, 12(8):307–312.

Gavrilets, S. (1999). A dynamical theory of speciation on holey adaptive landscapes. *The American Naturalist*, 154(1):1–22.

Gavrilets, S. (2000). Waiting time to parapatric speciation. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 267(1461):2483–2492.

Gavrilets, S. (2014). Models of Speciation: Where Are We Now? *Journal of Heredity*, 105(S1):743–755.

Gavrilets, S., Acton, R., and Gravner, J. (2000a). Dynamics of speciation and diversification in a metapopulation. *Evolution*, 54(5):1493–1501.

Gavrilets, S. and Gravner, J. (1997). Percolation on the fitness hypercube and the evolution of reproductive isolation. *J Theor Biol.*, 184(1):51–64.

Gavrilets, S., Hai, L., and Vose, M. D. (1998). Rapid parapatric speciation on holey adaptive landscapes. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 265(1405):1483–1489.

Gavrilets, S., Li, H., and Vose, M. D. (2000b). Patterns of parapatric speciation. *Evolution*, 54(14):1126–34.

Gordon, M. (1931). Hereditary basis of melanosis in hybrid fishes. *Amer. J. Cancer.*, 15:1495–1523.

Gourbiere, S. and Mallet, J. (2010). Are species real? The shape of the species boundary with exponential failure, reinforcement, and the 'missing snowball'. *Evolution*, 64(1):1–24.

Harvey, M. G., Seeholzer, G. F., Smith, B. T., Rabosky, D. L., Cuervo, A. M., and Brumfield, R. T. (2017). Positive association between population genetic differentiation and speciation rates in New World birds. *Proceedings of the National Academy of Sciences*, 114(24):6328–6333.

Harvey, M. G., Singhal, S., and Rabosky, D. L. (2019). Beyond reproductive isolation: Demographic controls on the speciation process. *Annual Review of Ecology, Evolution, and Systematics*, 50:75–95.

bioRxiv preprint doi: https://doi.org/10.1101/2023.11.09.564915; this version posted November 10, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Harvey, P. H., May, R. M., and Nee, S. (1994). Phylogenies without fossils. *Evolution*, 48(3):523–529.

Higgs, P. G. and Derrida, B. (1991). Stochastic models for species formation in evolving populations. *Journal of Physics A: Mathematical and General*, 24(17):L985.

Higgs, P. G. and Derrida, B. (1992). Genetic distance and species formation in evolving populations. *Journal of molecular evolution*, 35:454–465.

Hoelzer, G. A., Drewes, R., Meier, J., and Doursat, R. (2008). Isolation-by-distance and outbreeding depression are sufficient to drive parapatric speciation in the absence of environmental influences. *PLoS Computational Biology*, 4(7):e1000126.

Holley, R. A. and Liggett, T. M. (1975). Ergodic theorems for weakly interacting infinite systems and the voter model. *The Annals of Probability*, pages 643–663.

Hubbell, S. P. (2001). *The unified neutral theory of biodiversity and biogeography*, volume 32. Princeton University Press.

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(1):193–199.

Hubbell, S. P. and Lake, J. (2003). The neutral theory of biodiversity and biogeography, and beyond,[in:] tm blackburn & kj gaston (eds.), macroecology: patterns and process. *Blackwell, Oxford*, 45:63.

Irwin, D. E., Irwin, J. H., and Price, T. D. (2001). Ring species as bridges between microevolution and speciation. *Microevolution rate, pattern, process*, pages 223–243.

Jablonski, D. and Roy, K. (2003). Geographical range and speciation in fossil and living molluscs. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 270(1513):401–406.

Jabot, F. and Chave, J. (2009). Inferring the parameters of the neutral theory of biodiversity using phylogenetic information and implications for tropical forests. *Ecology Letters*, 12(3):239–248.

Kendall, D. G. (1948). On the generalized birth-and-death" process. *The Annals of Mathematical Statistics*, 19(1):1–15.

Kimura, M. and Crow, J. F. (1964). The number of alleles that can be maintained in a finite population. *Genetics*, 49(4):725.

Kondrashov, A. S. (2003). Accumulation of dobzhansky-muller incompatibilities within a spatially structured population. *Evolution*, 57(1):151–153.

Kopp, M. (2010). Speciation and the neutral theory of biodiversity: Modes of speciation affect patterns of biodiversity in neutral communities. *Bioessays*, 32(7):564–570.

Kuhner, M. K. and Yamato, J. (2015). Practical performance of tree comparison metrics. *Systematic Biology*, 64(2):205–214.

Lambert, A., Morlon, H., and Etienne, R. S. (2015). The reconstructed tree in the lineage-based model of protracted speciation. *Journal of Mathematical Biology*, 70(1):367–397.

Lambert, A. and Steel, M. (2013). Predicting the loss of phylogenetic diversity under non-stationary diversification models. *Journal of Theoretical Biology*, 337:111–124.

bioRxiv preprint doi: https://doi.org/10.1101/2023.11.09.564915; this version posted November 10, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Li, J., Huang, J.-P., Sukumaran, J., and Knowles, L. L. (2018). Microevolutionary processes impact macroevolutionary patterns. *BMC Evolutionary Biology*, 18(1):1–8.

Livingstone, K., Olofsson, P., Cochran, G., Dagilis, A., MacPherson, K., and Seitz Jr, K. A. (2012). A stochastic model for the development of bateson–dobzhansky–muller incompatibilities that incorporates protein interaction networks. *Mathematical biosciences*, 238(1):49–53.

Louca, S. and Pennell, M. W. (2020). Extant timetrees are consistent with a myriad of diversification histories. *Nature*, 580(7804):502–505.

MacArthur, R. and Wilson, E. (1967). *Island biogeography theory*. Princeton, NJ: Princeton University Press.

Maddison, W. P. (1997). Gene trees in species trees. *Systematic Biology*, 46(3):523–536.

Magurran, A. E. (2003). Measuring biological diversity. *Current Biology*, 31:R1174–R1177.

Mallet, J. (2007). Hybrid speciation. *Nature*, 446(7133):279–283.

Manceau, M. and Lambert, A. (2019). The species problem from the modeler's point of view. *Bulletin of Mathematical Biology*, 81(3):878–898.

Manceau, M., Lambert, A., and Morlon, H. (2015). Phylogenies support out-of-equilibrium models of biodiversity. *Ecology Letters*, 18(4):347–356.

Manzo, F. and Peliti, L. (1994). Geographic speciation in the Derrida-Higgs model of species formation. *J. Phys. A: Math. Gen.*, 27(7079).

Marin, J., Achaz, G., Crombach, A., and Lambert, A. (2020). The genomic view of diversification. *Journal of Evolutionary Biology*, 33(10):1387–1404.

Marques, D. A., Meier, J. I., and Seehausen, O. (2019). A combinatorial view on speciation and adaptive radiation. *Trends in Ecology & Evolution*, 34(6):531–544.

Martins, A. B., de Aguiar, M. A., and Bar-Yam, Y. (2013). Evolution and stability of ring species. *Proceedings of the National Academy of Sciences*, 110(13):5080–5084.

Matute, D. R. and Cooper, B. S. (2021). Comparative studies on speciation: 30 years since Coyne and Orr. *Evolution*.

Maya-Lastra, C. A. and Eaton, D. A. (2021 preprint). Genetic incompatibilities do not snowball in a demographic model of speciation. *bioRxiv*.

McGill, B. J., Etienne, R. S., Gray, J. S., Alonso, D., Anderson, M. J., Benecha, H. K., Dornelas, M., Enquist, B. J., Green, J. L., He, F., et al. (2007). Species abundance distributions: moving beyond single prediction theories to integration within an ecological framework. *Ecology letters*, 10(10):995–1015.

Miró Pina, V. and Schertzer, E. (2019). How does geographical distance translate into genetic distance? *Stochastic Processes and their Applications*, 129(10):3893–3921.

Moen, D. and Morlon, H. (2014). Why does diversification slow down? *Trends in Ecology & Evolution*, 29(4):190–197.

bioRxiv preprint doi: https://doi.org/10.1101/2023.11.09.564915; this version posted November 10, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Mooers, A., Gascuel, O., Stadler, T., Li, H., and Steel, M. (2012). Branch lengths on birth–death trees and the expected loss of phylogenetic diversity. *Systematic Biology*, 61(2):195–203.

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

Morlon, H., Robin, S., and Hartig, F. (2022). Studying speciation and extinction dynamics from phylogenies: addressing identifiability issues. *Trends in Ecology & Evolution*.

Muller, H. J. (1942). Isolating mechanisms, evolution and temperature. *Biol. Symp.*, 811:71–125.

Nee, S. C., May, R. M., and Harvey, P. H. (1994). The reconstructed evolutionary process. *Philos. Trans. Roy. Soc. London Ser. B*, 344:305–311.

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

O'Dwyer, J. P. and Chisholm, R. (2014). A mean field model for competition: from neutral ecology to the red queen. *Ecology letters*, 17(8):961–969.

O'Dwyer, J. P. and Green, J. L. (2010). Field theory for biogeography: a spatially explicit model for predicting patterns of biodiversity. *Ecology letters*, 13(1):87–95.

Orr, H. A. (1995). The population genetics of speciation: the evolution of hybrid incompatibilities. *Genetics*, 139(4):1805–1813.

Orr, H. A. (1996). Dobzhansky, bateson, and the genetics of speciation. *Genetics*, 144(4):1331.

Orr, H. A. and Orr, L. H. (1996). Waiting for speciation: the effect of population subdivision on the time to speciation. *Evolution*, 50(5):1742–1749.

Orr, H. A. and Turelli, M. (2001). The evolution of postzygotic isolation: accumulating dobzhansky-muller incompatibilities. *Evolution*, 55(6):1085–1094.

Patton, E. E., Mitchell, D. L., and Nairn, R. S. (2010). Genetic and environmental melanoma models in fish. *Pigment Cell and Melanoma Research*, 23(3):314–337.

Pennisi, E. (2016). Shaking up the tree of life. *Science*.

Pigot, A. L., Phillimore, A. B., Owens, I. P., and Orme, C. D. L. (2010). The shape and temporal dynamics of phylogenetic trees arising from geographic speciation. *Systematic biology*, 59(6):660–673.

Preston, F. W. (1948). The commonness, and rarity, of species. *Ecology*, 29(3):254–283.

Pybus, O. G. and Harvey, P. H. (2000). Testing macro–evolutionary models using incomplete molecular phylogenies. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 267(1459):2267–2272.

Pyron, R. A. and Burbrink, F. T. (2013). Phylogenetic estimates of speciation and extinction rates for testing ecological and evolutionary hypotheses. *Trends in ecology & evolution*, 28(12):729–736.

Rabosky, D. L. (2016). Reproductive isolation and the causes of speciation rate variation in nature. *Biological Journal of the Linnean Society*, 118(1):13–25.

bioRxiv preprint doi: https://doi.org/10.1101/2023.11.09.564915; this version posted November 10, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Rabosky, D. L. and Goldberg, E. E. (2015). Model inadequacy and mistaken inferences of trait-dependent speciation. *Systematic Biology*, 64(2):340–355.

Rabosky, D. L. and Matute, D. R. (2013). Macroevolutionary speciation rates are decoupled from the evolution of intrinsic reproductive isolation in Drosophila and birds. *Proceedings of the National Academy of Sciences*, 110(38):15354–15359.

Raup, D. M., Gould, S. J., Schopf, T. J., and Simberloff, D. S. (1973). Stochastic models of phylogeny and the evolution of diversity. *The Journal of Geology*, 81(5):525–542.

Ricklefs, R. E. (2007). Estimating diversification rates from phylogenetic information. *Trends in ecology & evolution*, 22(11):601–610.

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

Rosindell, J. and Cornell, S. J. (2007). Species-area relationships from a spatially explicit neutral model in an infinite landscape. *Ecol. Lett.*, 10(7):586–595.

Rosindell, J. and Cornell, S. J. (2009). Species–area curves, neutral models, and long-distance dispersal. *Ecology*, 90(7):1743–1750.

Rosindell, J., Cornell, S. J., Hubbell, S. P., and Etienne, R. S. (2010). Protracted speciation revitalizes the neutral theory of biodiversity. *Ecology Letters*, 13(6):716–727.

Roux, C., Fraisse, C., Romiguier, J., Anciaux, Y., Galtier, N., and Bierne, N. (2016). Shedding light on the grey zone of speciation along a continuum of genomic divergence. *PLoS Biology*, 14(12):e2000234.

Schweiger, O., Klotz, S., Durka, W., and Kühn, I. (2008). A comparative test of phylogenetic diversity indices. *Oecologia*, 157:485–495.

Seehausen, O., Butlin, R. K., Keller, I., Wagner, C. E., Boughman, J. W., Hohenlohe, P. A., Peichel, C. L., Saetre, G.-P., Bank, C., Brännström, Å., et al. (2014). Genomics and the origin of species. *Nature Reviews Genetics*, 15(3):176–192.

Shao, K.-T. and Sokal, R. R. (1990). Tree balance. *Systematic Zoology*, 39(3):266–276.

Singhal, S., Huang, H., Grundler, M. R., Marchán-Rivadeneira, M. R., Holmes, I., Title, P. O., Donnellan, S. C., and Rabosky, D. L. (2018). Does population structure predict the rate of speciation? A comparative test across Australia's most diverse vertebrate radiation. *The American Naturalist*, 192(4):432–447.

Stadler, T. (2013). Recovering speciation and extinction dynamics based on phylogenies. *Journal of Evolutionary Biology*, 26(6):1203–1219.

Tjørve, E. (2003). Shapes and functions of species–area curves: a review of possible models. *Journal of Biogeography*, 30(6):827–835.

Vacquier, V. D. and Swanson, W. J. (2011). Selection in the rapid evolution of gamete recognition proteins in marine invertebrates. *Cold Spring Harbor Perspectives in Biology*, 3(11).

Via, S. (2009). Natural selection in action during speciation. *Proceedings of the National Academy of Sciences*, 106(Supplement 1):9939–9946.

bioRxiv preprint doi: https://doi.org/10.1101/2023.11.09.564915; this version posted November 10, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Wagner, P. J. and Erwin, D. H. (1995). Phylogenetic patterns as tests of speciation models. In Erwin, D. H. and Anstey, R. L., editors, *New approaches to speciation in the fossil record*, pages 87–122. Columbia University Press.

Whittaker, R. H. (1960). Vegetation of the siskiyou mountains, oregon and california. *Ecological monographs*, 30(3):279–338.

Williams, C. B. et al. (1964). Patterns in the balance of nature and related problems of quantitative ecology. *Patterns in the balance of nature and related problems of quantitative ecology.*

Wolf, J. B. and Ellegren, H. (2017). Making sense of genomic islands of differentiation in light of speciation. *Nature Reviews Genetics*, 18(2):87.

Wu, C.-I. (2001). The genic view of the process of speciation. *Journal of Evolutionary Biology*, 14(6):851–865.

Yamaguchi, R. and Iwasa, Y. (2013). First passage time to allopatric speciation. *Interface Focus*, 3(6).

Yamaguchi, R. and Iwasa, Y. (2015). Smallness of the number of incompatibility loci can facilitate parapatric speciation. *Journal of Theoretical Biology*, 405:36–45.

Yamasaki, Y. Y., Takeshima, H., Kano, Y., Oseko, N., Suzuki, T., Nishida, M., and Watanabe, K. (2020). Ecosystem size predicts the probability of speciation in migratory freshwater fish. *Molecular Ecology*, 29(16):3071–3083.

Zillio, T., Volkov, I., Banavar, J. R., Hubbell, S. P., and Maritan, A. (2005). Spatial scaling in model plant communities. *Phys. Rev. Lett*, 95:098101.

bioRxiv preprint doi: https://doi.org/10.1101/2023.11.09.564915; this version posted November 10, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Appendices

A Appendix on the Red Queen model

In this appendix, we study the position and time of apparition of the first speciation event in the Red Queen model of Section 3.3.2. The goal is to give a sketch of the proof of the following theorem – focusing on the important ideas but leaving aside uninformative technicalities.

Theorem 1. *Let Z and T be, respectively, the position and time of the first speciation event in the continuous-space limit of the Red Queen model with mutation parameter μ. Then, as μ goes to 0, $(Z, \mu^2 T)$ converges in distribution to $(Z_{\lim}, T_{\lim})$, where*

(i) $Z_{\lim}$ a continuous random variable on $[0,1]$ with probability density function $f(z) = 6z(1-z)$;

(ii) $T_{\lim}$ is an exponential random variable with parameter $1/3$;

(iii) $Z_{\lim}$ and $T_{\lim}$ are independent.

Note that, in this theorem as in the rest of this section, the random variables Z and T depend on the parameter μ, but that we keep this dependence implicit for readability. We will also do so for other quantities.

Finally, throughout the rest of the section, we use the word *barrier* to refer to a point in $[0,1]$ that cannot be crossed after a speciation event (in other words, the place where to two mutants involved in the speciation event meet); see Figure 9.

A.1 Proof idea: limit of T

Let us start by introducing some notation. Let $P = \{(T_i, U_i)\}_{i\in\mathbb{N}^*}$ be the set of times and positions of apparition of the mutants, where T_i is the time of apparition of the i-th mutant and U_i is the point in $[0,1]$ where it appeared. Note that P is a Poisson point process on $\mathbb{R}_+ \times [0,1]$ with intensity measure $\mu \mathrm{d}t \otimes \mathrm{d}x$. Thus, $(U_i)_{i\in\mathbb{N}^*}$ is a sequence of independent and identically distributed (i.i.d.) uniform random variables on $[0,1]$, and there exists a sequence $(\xi_i)_{i\in\mathbb{N}^*}$ of i.i.d. exponential variables with parameter μ such that $(\xi_i)_{i\in\mathbb{N}^*}$ is independent of $(U_i)_{i\in\mathbb{N}^*}$ and, for all $i \in \mathbb{N}^*$,

$$T_i = \sum_{k=1}^{i} \xi_k.$$

Each mutant that appears at position x and time t will propagate to the left and to the right at speed 1. Therefore, it will encounter a younger mutant (i.e. an individual carrying a genotype that appeared after t) if and only if a mutant appears in the set $A_{x,t}$ corresponding to the yellow zone in Figure 11. Formally, for all $t \in \mathbb{R}_+, x \in [0,1]$, let

$$A^l_{x,t} = \{(s,y) \in \mathbb{R}_+ \times [0,1]; s \geq t, y \leq x + t - s\}$$

and

$$A^r_{x,t} = \{(s,y) \in \mathbb{R}_+ \times [0,1]; s \geq t, y \geq x + s - t\}$$

correspond to the left and right components of this set. Then, $A_{x,t} = A^l_{x,t} \cup A^r_{x,t}$. The area of $A_{x,t}$, which we denote by $\lambda(A_{x,t})$, is

$$\lambda(A_{x,t}) = \frac{x^2}{2} + \frac{(1-x)^2}{2}. \tag{4}$$

bioRxiv preprint doi: https://doi.org/10.1101/2023.11.09.564915; this version posted November 10, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

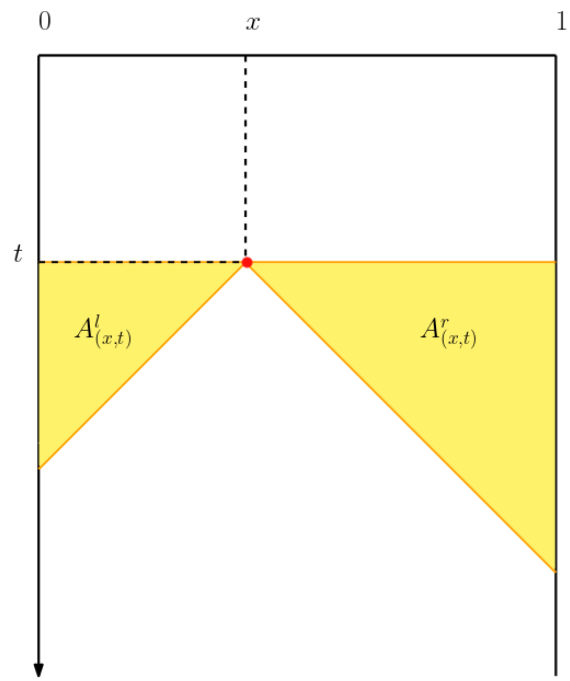


Figure 11: The mutation represented by a red point at (t, x) will create a barrier with a mutant that appeared after t if and only if at least one mutant appears in the yellow zone.

So far, we have assumed that the position where the first mutant appears is fixed, but in reality it is a random variable $(X, \theta) \in [0, 1] \times \mathbb{R}_+$, where X is uniform on $[0, 1]$. Therefore, $\lambda(A_{x,t})$ is the area of the zone where the apparition of a new mutant will create a barrier, "conditional on $(X, \theta) = (x, t)$" (here we use quotes because the probability that $(X, \theta) = (x, t)$ is equal to 0 for all (x, t), but this quantity is well-defined as a conditional expectation). Integrating against the law of X, we get that the (unconditional) expected area of this zone is

$$\int_0^1 \frac{x^2}{2} + \frac{(1-x)^2}{2}\,\mathrm{d}x = \frac{1}{3}. \tag{5}$$

Therefore, if we denote by I the index of the first mutant involved in a speciation event, and by T_I the time of apparition of that mutant, since all mutant appear at rate μ we have

$$I \sim \mathrm{Geom}\left(\frac{\mu}{3}\right),$$

i.e. I follows a geometric distribution with parameter 1/3. Moreover, since the $(\xi_i)_{i\geq 1}$ are exponentially distributed with parameter μ, an easy computation yields

$$T_I = \sum_{i=1}^{I} \xi_i \sim \mathrm{Exp}\left(\frac{\mu^2}{3}\right), \tag{6}$$

where $\mathrm{Exp}(\theta)$ denotes the exponential distribution with parameter θ and $\sim$ indicates equality in distribution.

Recall that T is the time of the first speciation event, that is to say the time when the I-th mutant and the other mutant involved in this speciation meet (in fact, as explained below, that other mutant is the $(I+1)$-th mutant with arbitrarily large probability as μ goes to 0; but that is not

bioRxiv preprint doi: https://doi.org/10.1101/2023.11.09.564915; this version posted November 10, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

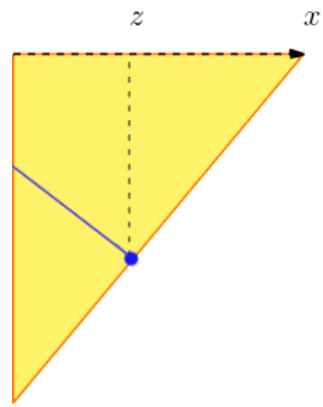


Figure 12: The blue line represents the set of points such that the apparition of a second mutant at one of these points will create a barrier at position z.

relevant for now). Since all mutants propagate at speed 1 in the interval $[0, 1]$, once the I-th mutant has appeared it cannot take more than one unit of time for the speciation to occur, irrespective of the initial positions of that mutant and of the other mutant involved in the speciation. Therefore,

$$T_I \leq T \leq T_I + 1. \tag{7}$$

Combining Equations (6) and (7) gives that, as μ goes to 0, $\mu^2 T$ converges in distribution to an exponential random variable with parameter $1/3$, proving point (i) of the theorem.

A.2 Proof idea: limit of Z

Keeping the notation of the previous section, recall that I is the index of apparition the first mutant involved in the first speciation event. When the mutation rate goes to 0, the probability that three mutants coexist between the apparition of the I-th mutant and the first speciation goes to 0. This entails that the other mutant involved in the first speciation event will have index $I + 1$ with probability that goes to 1 as μ goes to 0. Therefore, in the limit, the position of the first barrier is the point where the mutants I and $I + 1$ meet.

Note that if the I-th mutation appears at $x \in [0, 1]$, then the I-th and $(I + 1)$-th mutants meet at some point $z \in [0, x]$ if and only if the $(I + 1)$-th mutant appears somewhere in the blue segment depicted in Figure 12. The length L_z^l of this segment is

$$L_z^l = z\mathbb{1}_{\{0 \leq z \leq x/2\}} + (x - z)\mathbb{1}_{\{x/2 < z \leq x\}}.$$

Similarly, the set of potential points of apparition of the mutant $I+1$ that yield a barrier formed at $z \in]x, 1]$ has length

$$L_z^r = (z - x)\mathbb{1}_{\{x < z \leq (1+x)/2\}} + (1 - z)\mathbb{1}_{\{(1+x)/2 < z \leq 1\}}.$$

Thus, for any $z \in [0, 1]$, the probability density of Z_{lim} in z is proportional to $L_z = L_z^l + L_z^r$. Since

$$\int_0^1 L_x \, \mathrm{d}z = \frac{x^2 + (1 - x)^2}{2},$$

we get that, conditional on $\{X_I = x\}$, Z has density

$$z \mapsto \frac{2}{x^2 + (1 - x)^2} L_z \, .$$

bioRxiv preprint doi: https://doi.org/10.1101/2023.11.09.564915; this version posted November 10, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

To finish the proof, it suffices to determine the law of X_I and to integrate the conditional density of Z against it. The probability density of the position of apparition of a mutation that generates a speciation at position x is proportional to the area of the yellow zone in Figure 11, that is to say $\frac{x^2}{2} + \frac{(1-x)^2}{2}$. Therefore, the probability density of X_I is

$$x \mapsto \frac{\frac{x^2}{2} + \frac{(1-x)^2}{2}}{\int_0^1 \frac{u^2}{2} + \frac{(1-u)^2}{2} \mathrm{d}u} = \frac{3}{2}\left(x^2 + (1-x)^2\right),$$

and by integrating the conditional density of Z_{lim} against this density, we get that Z_{lim} has density $z \mapsto 6z(1-z)$, concluding the proof.