A Neutral Theory for Interpreting Correlations between Species and Genetic Diversity in Communities

Fabien Laroche,1,2,* Philippe Jarne,1 Thomas Lamy,1,3 Patrice David,1 and Francois Massol1,4

1. Centre d’Ecologie Fonctionnelle et Evolutive, Unité Mixte de Recherche (UMR) 5175, CNRS, Université de Montpellier, Université PaulValéry Montpellier, École Pratique des Hautes Etudes, 1919 Route de Mende, 34293 Montpellier, Cedex 5, France; 2. AgroParisTech, 19 Avenue du Maine, 75732 Paris, Cedex 15, France; 3. Département de Sciences Biologiques, Université de Montréal, Montréal, Québec, Canada; 4. Laboratoire de Génétique et Evolution des Populations Végétales, UMR 8198, CNRS, Université Lille 1, Bâtiment SN2, 59655 Villeneuve d’Ascq Cedex, France

Submitted March 4, 2014; Accepted July 19, 2014; Electronically published December 16, 2014

Online enhancements: appendixes, data file.

abstract: Spatial patterns of biological diversity have been extensively studied in ecology and population genetics, because they reflect the forces acting on biodiversity. A growing number of studies have found that genetic (within-species) and species diversity can be correlated in space (the so-called species-gene diversity correlation [SGDC]), which suggests that they are controlled by nonindependent processes. Positive SGDCs are generally assumed to arise from parallel responses of genetic and species diversity to variation in site size and connectivity. However, this argument implicitly assumes a neutral model that has yet to be developed. Here, we build such a model to predict SGDC in a metacommunity. We describe how SGDC emerges from competition within sites and variation in connectivity and carrying capacity among sites. We then introduce the formerly ignored mutation process, which affects genetic but not species diversity. When mutation rate is low, our model confirms that variation in the number of migrants among sites creates positive SGDCs. However, when considering high mutation rates, interactions between mutation, migration, and competition can produce negative SGDCs. Neutral processes thus do not always contribute positively to SGDCs. Our approach provides empirical guidelines for interpreting these novel patterns in natura with respect to evolutionary and ecological forces shaping metacommunities.

Keywords: neutral theory, SGDC, coalescence, community genetics, diversity pattern, mainland-island model.

Introduction

It has been recognized for several decades that the diversity patterns of genes within species and those of species within communities are not independent. Understanding their interactions is the main goal of “community genetics” (Antonovics 1976), an interdisciplinary field that has recently seen a burst of interest (Wares 2002; Agrawal 2003; Neuhäuser et al. 2003; Bernhardsson et al. 2013). The rise of environmental genomics and long-term surveys of populations and communities has enhanced the opportunity to confront these two organizational levels (Gugerli et al. 2013). In particular, it is becoming common practice to compute “species-genes diversity correlation” (SGDC; Vellend 2003), which consists in quantifying the link between the genetic diversity in several populations of a species (the focal species) and the species diversity of the local communities within which these populations are embedded. This has been done in more than 40 studies since the seminal work of Vellend and coworkers (see Vellend et al. 2014 for a review).

SGDCs provide information on both fundamental and applied issues with regard to biodiversity. From a fundamental perspective, investigating the generality of SGDC patterns can help to uncover determinants of ecological processes shaping diversity at different levels. For example, several empirical studies have shown that site area (Vellend 2003), which often constitutes a proxy for carrying capacity in community ecology, and site connectivity (Lamy et al. 2013) contribute markedly to positive SGDCs. This suggests that drift and migration might have a strong impact on both species and genetic diversity. From a more applied perspective, detecting positive SGDCs might be useful, in conservation studies, to infer diversity from one level to the other (e.g., predicting species diversity based on genetic data; Papadopoulou et al. 2011).

The growing number of empirical studies on SGDCs constitute a strong incentive to build a quantitative theoretical basis that would help interpreting observed patterns. Vellend and Geber (2005) made a conceptual advance on this issue by envisioning three types of relationships between diversity levels that may generate interpretable signals in community genetics: causal effects of genetic diversity on species diversity, causal effects of...
species diversity on genetic diversity, and simultaneously parallel effects of external factors on both levels. Neutral theories of molecular evolution (Kimura 1984) and of biodiversity (Hubbell 2001) provide some conceptual elements regarding these potential parallel effects. Indeed, both theories consider limited dispersal and drift to be the main drivers of diversity patterns, and they both predict that carrying capacities and immigration rates should be positively related to diversity (Wright 1931; Hubbell 2001). A positive SGDC should then arise from any external factor generating variation in carrying capacity and connectivity across sites, as has been supported by simulation work (Vellend 2005).

However, even under a neutral framework, the interpretation of positive SGDCs may not be as straightforward as suggested above, because of interactions between the focal species (i.e., the one studied for genetic variation) and other species of the community within sites. In particular, the local abundance of a species may be positively linked to its genetic diversity but also negatively linked to the abundance of other species, and thus to species diversity, as a consequence of limited space. This might produce a negative SGDC (Vellend 2005; Wehenkel et al. 2006; Odat et al. 2010) under specific circumstances that remain to be characterized quantitatively. To our knowledge, no analytical model predicts the sign and magnitude of SGDCs when accounting for the two effects mentioned above, namely, (i) local competition dynamics and (ii) variation in carrying capacity and connectivity among sites. A first objective here is to propose such a model.

A complete quantitative theory of SGDCs also has to include the forces generating diversity, namely, mutation and speciation. These processes have indeed been neglected when discussing SGDC on the grounds that they are often too slow compared with ecological processes (Vellend and Geber 2005). This is true when these rates are negligible with respect to migration. For speciation, this assumption may be challenged when considering archipelagoes (Losos and Schluter 2000) but remains correct when studying metacommunities at limited spatial scale (e.g., a pond network in a single island; Lamy et al. 2013). Here we focus on situations where speciation can be neglected. However, even in this context, assuming that mutation has a negligible impact on genetic diversity is still questionable, especially when using highly mutable markers such as microsatellites (Jarne and Lagoda 1996; Ellegren 2002). Such markers are commonly used in studies reporting SGDCs (Cleary et al. 2006; He et al. 2008; Struebig et al. 2011; Blum et al. 2012; Lamy et al. 2013). Our second objective is thus to provide insights on how mutation may affect SGDCs, even at rather limited spatial and temporal scales.

We build a spatially implicit model of a metacommunity using a unifying neutral framework for both genetic and species dynamics to generate theoretical expectations on SGDCs. Our approach takes into account drift and migration at both diversity levels, as well as mutation, while speciation is neglected. We consider a set of local communities receiving migrants from a larger regional community (Hubbell 2001). This model allows distinguishing within- and among-site effects on SGDCs and thus disentangling the effects of competition within local sites from those of drift and migration among sites. When mutation is neglected, the SGDC turns out to be positive. However, high mutation rates, compared with immigration rates, can produce negative SGDCs. Even under neutral assumptions, the sign of SGDCs can be labile, and understanding SGDCs is therefore not straightforward. On the basis of our framework, we provide some empirical guidelines for interpreting SGDCs.

**Material and Methods**

**Modeling the Dynamics of Species and Gene Diversity in a Site**

Our work is based on an individual-based model derived from the classical neutral model of ecological communities (Hubbell 2001). We describe it hereafter following the standardized “overview, design concepts, and details” protocol (Grimm et al. 2010).

**Purpose.** The model aims at simultaneously providing the species composition of a sample taken from a community and the genotypes of the individuals that belong to the focal species in this sample. Model predictions are based on two features of the sampled site: the carrying capacity (K) and the immigration rate from the regional pool (m).

The symbols used are summarized in table 1.

**Entities, State Variables, and Scale.** The model contains three types of entities: a site, its individuals, and a regional pool serving as a source of migrants. Individuals are described using two state variables: the species they belong to and, for individuals that belong to the focal species, their allelic state at a given locus (under the assumption of haplody). The latter variable is ignored for individuals that do not belong to the focal species. The site is described by parameters K and m (which are permanent characteristics) and the list of individuals that it contains (which varies in time). The regional pool of individuals is characterized by a set of constant state variables, including the relative abundances of B species \( f = (f_1, f_2, \ldots, f_B) \) and a parameter \( \theta \) which quantifies the mutation-drift ratio in the regional population of the focal species (app. A; apps.
A–C available online). When not neglected, mutation is characterized by a per-birth mutation rate \( \mu \) in the focal species.

**Process Overview.** The model is characterized by discrete death-birth cycles in the site. At the beginning of each cycle, the site contains exactly \( K \) individuals (i.e., is saturated). An individual is then randomly chosen, discarded, and replaced by the offspring of a reproducer which either belongs to the site, with probability \( 1 - m \), or to the regional pool, with probability \( m \). When the reproducer belongs to the site, the offspring inherits its species identity. Its genotype (focal species) is either the same as the reproducer’s genotype (with probability \( 1 - \mu \)) or a mutated allele not already present in the species (with probability \( \mu \); see below for additional discussion of the mutation regime). When the reproducer belongs to the regional pool, the species identity of the immigrant offspring is randomly drawn from the distribution of the species relative abundances in the regional pool \( f \). The offspring belongs to the focal species (with probability \( f \)), its genotype is determined as explained below. Note that, in our model, competition among genotypes and species occurs during these cycles, when dead individuals are replaced by offspring of either migrant or local origin (i.e., a lottery competition for space).

Two scenarios are considered with respect to mutation. The first scenario corresponds to a weak mutation regime \(( \mu \ll m \); in practice, \( \mu \) is set to zero) in which mutation is neglected in the local community dynamics. At the regional scale, the allelic frequencies of the focal species are assumed to be at mutation-drift equilibrium and follow a Ewens distribution with parameter \( \theta \) (Ewens and Tavare 2006). The second scenario corresponds to a strong mutation regime where mutation at the focal locus cannot be neglected when compared to migration \( \left( \mu \approx m \right) \). Mutation process follows an infinite-allele model: any mutation event generates an allele that never occurred before in the site. As a consequence of high mutation rate, the regional allelic pool is assumed to be infinitely diverse (app. A): immigrants always harbor alleles that did not occur before in the site.

**Outputs.** Species diversity and allelic diversity are determined through a sampling process designed to mimic a typical SGDC study. \( S \) individuals are randomly sampled from the site (the species sample; fig. 1). The species composition of this sample is described by \( s = (s_1, s_2, \ldots, s_p) \), where \( s_i \) individuals belong to species \( i \), and \( \sum_i s_i = S \). Species diversity is computed as species richness \( R_{\text{spe}} \); i.e., as the number of distinct species occurring in \( s \). In the species sample, the allelic states of the individuals belonging to the focal species \( e \) are described by \( t = (t_1, t_2, \ldots, t_r) \), where \( t_i \) individuals carry allele \( j_i \), and
Figure 1: Sampling protocol of a site in the model. The large rectangle on the left depicts a site. Arrows represent random sampling. Dashed rectangles represent samples, and pictograms represent species. The focal species (squares) harbors alleles that are depicted with different graphical patterns (crosses and points). The genetic sample u is obtained by subsampling k individuals among individuals of the focal species (t) included in the species sample (s). Here S = 7 and k = 2.

Σj t, j = s. A random subset, u, of t containing k individuals is genotyped and constitutes the “genetic sample” within sites. The genetic diversity is estimated using allelic richness (R_all), computed as the number of distinct alleles occurring in u. Note that with this sampling procedure, allelic and species richness can be computed only for sites containing more than S individuals (K > S) and yielding a sample s containing more than k individuals of species e (s, > k).

Modeling and Decomposing Species-Gene Diversity Relationships across Sites

The influence of variation in carrying capacity (K) and immigration rate (m) among sites on SGDC is modeled by considering a set of sites created by independently drawing values of K and m from a bivariate distribution with given variance and covariance (app. B). All the sites are connected to the same regional pool and follow the same mutation dynamics (i.e., weak or strong). R_all and R_SPE are computed using our model in all the sites where K > S and s, > k (see above). Note that our sampling protocol controls for sample size at both species (through S) and genetic levels (through k), which allows comparing diversity measures among sites. We provide below theoretical expectations about the sign of the expected covariance between R_all and R_SPE computed across sites (Csg). Although Csg is not the SGDC classically estimated in empirical studies (i.e., authors generally use Pearson’s correlation coefficient), it provides qualitative information about the sign of the expected relationship between genetic and species diversity. Besides, Csg can be decomposed into two effects. The first one occurs within sites as the result of local competition. The second effect stems from the variation in carrying capacity (K) and migration (m) among sites. Technically speaking, this can be expressed as the decomposition of Csg as the sum of two covariances, C_within and C_among (app. C), with

\[
\begin{align*}
C_{sg} &= C_{within} + C_{among} \\
C_{within} &= E[Cov_{K, m}[R_{spe}, R_{all}]] \\
C_{among} &= \text{Cov}[R_{spe}(K, m), R_{all}(K, m)]
\end{align*}
\]

where Cov_K, m [R_{spe}, R_{all}] is the covariance between specific and allelic richness (considered as random variables) within a site with given K and m values, E, and Cov are the expectation and covariance over (K, m) distribution, respectively, and overlined quantities are expectations within sites with given K and m values. C_among reflects the effect of (K, m) variation among sites. Importantly, C_among is null when K and m do not vary among sites, in which case only local competition (C_within) determines Csg and thus the sign of SGDC. From a statistical point of view, this decomposition of Csg can be interpreted as in an analysis of variance framework, C_among being the part of covariance explained by K and m and C_within being the residual covariance.

Simulating R_all and R_SPE in a Set of Local Sites

Simulations illustrate our theoretical predictions about the sign of SGDC and provide more quantitative information about SGDC (i.e., Pearson’s correlation coefficient) variation with respect to K and m distribution among sites. An efficient sampling approach in our model is to simulate the genealogy of the S individuals per sample backward in time (coalescence approach; Rosindell et al. 2008). This simulation strategy is used here to generate s and u samples within local sites, from which R_all and R_SPE are computed. More details about the simulation algorithm are provided in supporting material.

In all the simulations, the regional community contains 20 species, the relative abundances of which are derived from a truncated geometric distribution with parameter 0.2 (i.e., \(f_i = (1 - 0.2) \times 0.2^{-1}/(1 - 0.2^m))\). The most abundant species in the regional pool is chosen as the focal one (\(f_i \approx 0.8\)) to avoid discarding many sites because of unsuccessful sampling; this is a reasonable assumption with regard to empirical studies reporting SGDCs, which generally analyze genetic diversity in common species. Under weak mutation, \(\theta\) is set to 10. Under strong mutation, \(\mu\) is set to \(10^{-3}\), in line with what is known for microsatellite markers (Jarne and Lagoda 1996; Ellegren 2002). Landscapes considered here are sets of 100 sites. K and m per site are determined by sampling (log(K – 1), log(m/(1 – m))) in a “discretized” bivariate Gaussian distribution with mean \((\alpha_K, \alpha_m)\), marginal variances
expression of in equation (1) can therefore be re-
relative strength of drift and immigration within sites. The
Olff 2004; Etienne and Alonso 2005), which quantifies the
called effective number of migrants (app. A; Etienne and
pends on the variation in ( among sites. It turns out
serves a proxy for the abundance of the focal species within
This result is a consequence of controlling for genetic sam-
composition of the covariance between diversity levels
An important outcome of our work is to provide a de-
composition of the covariance between diversity levels
(C_all) into effects occurring within (C_within) and among sites
(C_among) and derive predictions about each of them. We
separately analyzed each type of effects in our model and
generated conclusions about the overall sign and value of
SGDC. Because the mutation-to-migration ratio strongly
generated conclusions about the overall sign and value of
SGDC. Because the mutation-to-migration ratio strongly
and increase with
I. The variance in I is positively
related to the variances in both K and m as well as to the
covariance between K and m. Note, however, that a large
variance in both K and m, associated to a strong negative
covariance between these two parameters, generates a low
variance in I, leading to weak values of C_all and SGDCs
(app. B).

Results

An important outcome of our work is to provide a de-
composition of the covariance between diversity levels
(C_all) into effects occurring within (C_within) and among sites
(C_among) and derive predictions about each of them. We
separately analyzed each type of effects in our model and
generated conclusions about the overall sign and value of
SGDC. Because the mutation-to-migration ratio strongly
and increase with
I. The variance in I is positively
related to the variances in both K and m as well as to the
covariance between K and m. Note, however, that a large
variance in both K and m, associated to a strong negative
covariance between these two parameters, generates a low
variance in I, leading to weak values of C_all and SGDCs
(app. B).

Weak Mutation Regime

The behavior of the within- and among-site components of
C_mig can be analyzed by considering the joint probability of
s and u. Under the weak mutation regime, we estab-
lished that the compositions of s and u are probabilistically
independent within a site (app. A), so that C_within is null.
This result is a consequence of controlling for genetic sample
size (through the parameter k here) when estimating
genetic diversity, thus dampening any effect of local pop-
ulation size on R_all. We performed repeated simulations of
a single site with given K and m values to estimate the
relation between genetic and species diversity and the
abundance of the focal species in the site. Because s provides
a proxy for the abundance of the focal species within
sites, species and genetic diversity were actually sorted as
a function of s (fig. 2). As predicted by our theoretical
analysis (see app. A), R_all decreases with s, but R_all does
not show any trend with respect to s.

As the within-site component is null, the covariance
between species and genetic diversity under the weak mu-
tation regime reduces to the C_among component, which de-
pends on the variation in (K, m) among sites. It turns out
That the latter influences s and u compositions only
through the variation in I = (K − 1)m/(1 − m), the so-
called effective number of migrants (app. A; Etienne
and Olff 2004; Etienne and Alonso 2005), which quantifies the
relative strength of drift and immigration within sites. The
expression of C_among in equation (1) can therefore be re-
written as

\[ C_{among} = \text{Cov} \left[ R_{spe}(I), R_{all}(I) \right], \]

where overlined quantities are expectations in sites with
parameter I. It can be shown (app. A) that both R_spe(I)
and R_all(I) increase with I, so that C_among is expected to be
positive.

From these results, C_all (equal to C_among) always takes
positive values. Simulations (fig. 3) illustrate these theo-
etical expectations: simulated SGDCs are always positive.
Moreover, SGDCs increase with the variance in I and, for
a given variance in I, they show very little variation. These
results are consistent with our theoretical prediction: both
R_spe and R_all depend on site parameters through the value
of I only and increase with I. The variance in I is positively
related to the variances in both K and m as well as to the
covariance between K and m. Note, however, that a large
variance in both K and m, associated to a strong negative
covariance between these two parameters, generates a low
variance in I, leading to weak values of C_all and SGDCs
(app. B).

Figure 2: Variation in mean R_spe and R_all as a function of s (sampling size of the focal species) in a site with given carrying capacity (K) and connectivity (m). A total of 10,000 simulations were performed under each mutation regime. Simulation outputs are sorted according to s value, and for each s value, the mean values of observed R_all (black circles), R_all under weak mutation (gray squares), and R_all under strong mutation (gray triangles) are reported. For R_all, the output of the 20,000 simulations are considered together to compute mean values for each s (because R_all does not depend on the mutation regime). A 95% confidence interval (1.96 times the standard error) is given with the R_all and R_all mean values. Other parameters are set to K = 1,000, m = 0.001, B = 20, f_i = (1 − 0.2) × 0.2^2/(1 − 0.2^2), e = 1, f_i = 0.8, θ = 10 (for weak mutation), μ = 10^−3 (for strong mutation), S = 50, and k = 5.
Under the strong mutation regime, is not necessarily landscapes that share the same values. However, detailed in figure 4 should also occur in more complex homogeneous landscapes (across sites. Thus the negative impact observed in the order 0 on in equation (1) yields the following and their immigration rate is small. A delta method to approximation:

\[
C_{\text{within}} \approx \text{Cov} \left( \text{spe} \{I, K \}, \text{all} \{I, m \} \right)
\]

\[
\approx \frac{\partial R_{\text{spe}}}{\partial I} \times \frac{\partial R_{\text{all}}}{\partial I} \text{Var} (I)
\]

\[
+ \frac{\partial R_{\text{spe}}}{\partial I} \times \frac{\partial R_{\text{all}}}{\partial m} \text{Cov} [I, m],
\]

where the notation is identical to that in equation (2), \(\partial\) is the symbol for partial derivative, and the approximation

\[
C_{\text{within}} \approx \text{Cov} \left( \left( \text{spe} \{I, K \}, \text{all} \{I, m \} \right) \right),
\]

where \(\overline{K}, \overline{m}\) are means of \(K\) and \(m\) across the landscape and \(\text{floor()}\) is the integer part operator. Other notations are similar to equation (1). We expect \(C_{\text{within}}\) to depend mostly on mean carrying capacity and immigration in the set of sites and not on variance and covariance of \(K\) and \(m\) across sites. Thus the negative impact observed in the homogeneous landscapes \((K\) and \(m\) constant across sites) detailed in figure 4 should also occur in more complex landscapes that share the same \((\overline{K}, \overline{m})\) values. However, the total covariance will result from the addition of \(C_{\text{within}}\) and \(C_{\text{among}}\), which may have different signs.

Mutation also has an impact on \(C_{\text{among}}\). As under the weak mutation regime, \(C_{\text{among}}\) is affected by variation in \(I\), but also by variation in migration alone \((m)\), independently from \(I\). This is because, when migration is high, mutation events have less impact on within-site diversity, which depends mainly on new alleles brought by immigrants (app. A). The expression of \(C_{\text{among}}\) is more complex than under the weak mutation regime (eq. [2]), because genetic diversity depends on both \(I\) and \(m\) as follows:

\[
C_{\text{among}} = \text{Cov} [\text{spe} \{I, K \}, \text{all} \{I, m \}]
\]

\[
= \frac{\partial R_{\text{spe}}}{\partial I} \times \frac{\partial R_{\text{all}}}{\partial I} \text{Var} (I)
\]

\[
+ \frac{\partial R_{\text{spe}}}{\partial I} \times \frac{\partial R_{\text{all}}}{\partial m} \text{Cov} [I, m],
\]

Under the strong mutation regime, \(C_{\text{within}}\) is not necessarily zero anymore. Indeed, \(R_{\text{all}}\) increases with \(s_{m}\) whereas \(R_{\text{spe}}\) tends to decrease (fig. 2), generating negative expectations for \(C_{\text{within}}\). This clearly appears when simulating homogeneous landscapes, with the same \((K, m)\) values in all sites (i.e., \(C_{\text{among}} = 0\); fig. 4). The SGDC is negative, especially when the carrying capacity \(K\) of sites is large and their immigration rate \(m\) is small.

A delta method to order 0 on \(C_{\text{within}}\) in equation (1) yields the following approximation:

\[
C_{\text{within}} \approx \text{Cov} \left( \text{spe} \left( \text{floor} (K), m \right), \text{all} \left( R_{\text{spe}}, R_{\text{all}} \right) \right),
\]

where \(R_{\text{spe}}, R_{\text{all}}\) are set to \(0\), \(2\), \(3\), \(3\), \(3\), \(1\), \(1\), \(\theta = 10\), \(S = 50\), and \(k = 5\).
Community genetics is a rising field of research that has developed along several lines, such as studying relation-
An important result of our work is that the covariance between species richness and allelic richness ($C_{\text{ST}}$) can be additively decomposed into (i) the effect of competition between the focal species and other species within local sites (the $C_{\text{within}}$ term) and (ii) the parallel effect of variation in carrying capacity and immigration rate of sites on allelic and species richness (the $C_{\text{among}}$ term; eq. [1]). Both effects had previously been identified in the literature. Local competition was thought to negatively affect SGDCs (i.e., $C_{\text{within}} < 0$ in our framework; Vellend 2005; Wehenkel et al. 2006), while variation in area and isolation among sites was thought to generate positive SGDCs (i.e., $C_{\text{among}} > 0$; Vellend 2003). The dominant effect could, in principle, be inferred from the SGDC sign without further need of a quantitative framework. However, our predictions are partially at variance with these intuitions. When the mutation rate is much lower than the immigration rate, $C_{\text{among}}$ is positive, as expected, but $C_{\text{within}}$ is always zero. When the mutation rate is comparable to or higher than the immigration rate, $C_{\text{within}}$ is negative, which corresponds to expectations, but $C_{\text{among}}$ can take both signs, which does not.

Because mutation drastically changes the predictions on SGDC patterns, an important aspect in empirical studies should be to determine the mutation-to-migration ratio before interpreting SGDCs. In particular, many estimates of SGDCs are based on microsatellites when evaluating the genetic diversity (Cleary et al. 2006; He et al. 2008; Struweig et al. 2011; Blum et al. 2012; Lamy et al. 2013). These markers may have high mutation rates (Jarne and Lagoda 1996; Ellegren 2002) that are potentially high enough to compare with immigration rates, especially in isolated sites. Insights on the mutation-to-migration ratio can be obtained by computing the relationships between a proxy of the number of migrants in a site ($f$), a proxy of the local carrying capacity ($K$), and the allelic richness of the focal species. If allelic richness increases with carrying capacity when controlling for the number of migrants (which can be assessed with a partial correlation analysis for instance), this suggests that mutation contributes strongly to the build-up of variation in these sites. Alternatively, one can also directly use the genetic polymorphism of the focal species to assess the relative strength of mutation and migration processes. For instance, when considering microsatellites, testing for a significant difference between $\hat{K}_{\text{ST}}$ and $F_{\text{ST}}$ estimators of genetic structure in the focal species could help in evaluating whether mutation could be neglected (Hardy et al. 2003). We note that speciation may have an impact on SGDC patterns similar to that of mutation. Speciation was not considered in our model, because we focused on a temporal/spatial scale at which it is unlikely to generate species variation to a significant extent. When the immigration and speciation rates are of

Figure 6: Values of expected species-gene diversity correlations (SGDCs) with respect to the variation in $I$ and the covariance between $m$ and $I$ across sites (i.e., the $[\eta^2, \rho_{\text{ST}}]$ space) for simulated landscapes under the strong mutation regime. A total of 500 simulations were performed for each landscape, and mean SGDC was represented here by a dot, with shades of gray indicating the associated value. The parameters were set to $\alpha_s = 3, \alpha_m = -3, \sigma_s^2 = 3, \sigma_m^2$ and $\rho_{\text{ST}}$ were numerically explored in $[0, 1] \times [-1, 1]$ by step of 0.01, leading to 19,900 combinations. Other parameters were set to $B = 20, f_s = (1 - 0.2) \times 0.2^{0.2}/(1 - 0.2)^{0.2}, e = 1, f_1 \approx 0.8, \mu = 10^{-5}, S = 50,$ and $k = 5$. 

ships between herbivore communities and the genetics of plants supporting these communities at small geographic scale (Bernhardsson et al. 2013; McArt and Thaler 2013) or using phylogeography to better understand past processes that have shaped present communities at larger scale (Wares 2002; Webb et al. 2002). The study of correlations between genetic diversity at the species level and species diversity at the community level (SGDCs; Vellend and Geber 2005) is one of these offshoots. Up to now, predictions on SGDCs have essentially been formalized on the basis of verbal models. We propose here a theoretical framework encompassing both species and genetic levels to more fully analyze and interpret SGDCs. We based our work on geological approaches and sampling formulae that are now commonly used in both population genetics (Wakeley 2008) and community ecology (Etienne and Olff 2004) to infer processes from patterns. Those techniques proved very useful, for instance, when searching for selective processes (Fu and Li 1993; Etienne 2005, 2007). Although sampling formulae are available for modeling both genetic and species dynamics (Etienne and Alonso 2005), we know of no previous work generating simultaneous predictions at both levels on the basis of a generalized coalescent of genes and species.
similar magnitude (e.g., in isolated metacommunities, such as archipelagos), a larger number of endemic species should be generated in sites with lower immigration rates (Rosindell and Phillimore 2011). Interactions between the effects of speciation and connectivity on SGDC should then be similar to those detected in our model about mutation.

When mutation is weak compared with immigration, our model predicts that local competition should not affect the SGDC pattern. We emphasize here the importance of the sampling protocol. Earlier studies predicted a negative $C_{\text{within}}$, because genetic diversity of the focal species is expected to increase with its population size (Vellend 2005; Wehenkel et al. 2006). The sampling protocol of our model (fig. 3) did not allow this positive relationship to occur, because a fixed number of individuals of the focal species were genotyped ($k$ individuals in sample $r$; smaller samples were disregarded). Other sampling protocols can generate the same disconnection between the allelic richness and the population size of the focal species as long as they incorporate a control of the genetic sample size. For instance, the same disconnection occurs when genetic diversity is measured by genotyping all individuals belonging to the focal species ($t$) and computing a rarefied richness indicator (Petit et al. 1998), as most empiricists do. Under weak mutation, controlling for the size of species and genetic samples filters out the influence of local competition, which facilitates the interpretation of observed patterns.

As the among-sites effect is always positive under weak mutation, our neutral theory yields the simple prediction that SGDC should always be positive and reflects the variance of the effective number of migrants among sites (fig. 3). Empirical studies that (i) demonstrated that mutation is weak, (ii) controlled for sample size in the sampling protocol, and (iii) observed strong variation in size and connectivity among sites should then expect a positive SGDC. When this prediction is not verified, it may mean one of three things. First, variation in size and connectivity may be negatively correlated among sites in such a way that the overall variation in the number of immigrants among sites is low (app. B). Second, there may be a non-neutral process at work. For instance, Derry et al. (2009) illustrated how species-sorting along an environmental gradient may contribute to cancel the positive parallel effects of variation in size and connectivity on SGDC. Finally, some other assumptions of our model may be violated. The last explanation may apply when considering spatially continuous systems (e.g., alpine forest; Taberlet et al. 2012) for which our implicit description of space may prove insufficient to describe the spatial autocorrelation in the system.

Under strong mutation, the analysis of SGDC patterns is different, because the correlation sign predicted by the neutral theory developed here is more labile than under the weak mutation regime. On the one hand, local competition ($C_{\text{within}}$ term) has a negative impact on SGDC. Indeed, the positive relationship between the population size of the focal species and the allelic richness of the genetic sample is maintained, and this occurs even when controlling for genetic sample size (fig. 3). On the other hand, $C_{\text{among}}$ can take either sign depending on the co-distribution of carrying capacities and immigration rates among sites. In particular, a negative $C_{\text{among}}$ value emerges when sites tend to receive the same effective number of immigrants per generation irrespective of their carrying capacity (i.e., low variance in $t$). Such a situation could occur, for example, in fragmented landscapes with patches of different sizes connected by corridors; the effective number of immigrants would primarily depend on the presence of corridors and may be uncorrelated to patch size (which determines its carrying capacity). On the whole, any sign of the SGDC is compatible with the neutral framework when mutation is strong, so that, contrary to the weak mutation regime, neutrality cannot be rejected on the basis of the sign of the correlation only. Note, however, that using markers that show different levels of polymorphism (using polymorphism as a proxy of mutation rate) may provide further tests. If the correlation is positive when using poorly variable markers and negative when considering highly variable ones, the overall observation is compatible with the neutral framework. By contrast, a consistently negative SGDC, whatever the level of polymorphism of the considered marker, may be interpreted as a rejection of our neutral model. When our framework applies, interpreting the SGDC sign under strong mutation is not straightforward. A significantly positive SGDC indicates a strong positive $C_{\text{among}}$ and can be interpreted as an effect of high variance in the number of migrants among sites. However, nonsignificant and negative SGDCs lead to ambiguous interpretation. In particular, negative correlations can indicate an effect of local competition but can also result from a negative $C_{\text{among}}$.

One way to progress in the interpretation of SGDCs is to decompose the covariance between species and genetic diversity ($C_{sq}$) into the $C_{\text{within}}$ and $C_{\text{among}}$ effects. Some authors suggested statistical methods to analyze the contribution of size and connectivity of sites to the overall SGDC (Vellend 2003; Lamy et al. 2013). Both studies detected significantly positive SGDCs along with a strong contribution of area and connectivity, respectively, to these correlations, which may indicate a strong positive $C_{\text{among}}$. Our model provides a theoretical basis for going one step further in this analysis by allowing a covariance decomposition based on mechanisms (instead of environmental factors) to be performed in empirical studies. One approach could be to estimate $I$ in sites and to directly per-
form covariance decomposition along those estimates instead of using proxies of size and connectivity, as done in former studies. This should provide a more direct assessment of $C_{\text{among}}$. However, estimating $I$ is not straightforward. One solution could be to use loci different from those used to compute SGDCs and to independently assess the migration-drift ratio in populations of the focal species through $Nm$ ($N$ and $m$ are the population size and immigration rate, respectively, in, for example, island models of population structure; Rousset 2004), which should provide a relevant proxy for $I_f$. Separate estimates of $f$ could be obtained by other approaches, such as by pooling all the local species samples to generate a regional sample (Jabot et al. 2008) so that $I$ could be isolated. Ultimately, decomposing SGDC patterns should contribute to a deeper understanding than the sign of SGDC alone. Beyond helping to interpret ambiguous cases such as negative SGDCs under strong mutation, disentangling $C_{\text{within}}$ and $C_{\text{among}}$ may also provide new tests of our framework: for instance, under low mutation, observing a large positive SGDC but no significant $C_{\text{among}}$ may indicate other non-neutral processes, such as positive interactions between the focal species and the rest of the community within sites (e.g., facilitation in plant communities; Brooker et al. 2008).

With the building of an adequate theory, SGDC patterns may be used to study the processes, such as dispersal and drift, acting in metacommunities. Certainly, a further step is the development of neutral models, including a full sampling theory to provide useful null hypotheses to detect selective processes, based on both species count data and genomic sequencing. The spectacular increase in the availability of genomic data opens interesting perspectives. It seems unlikely, however, that comparison of local diversity across levels provides enough information to unravel the complex processes acting in metacommunities, such as niche structure and environmental filtering among sites. Interestingly enough, empirical studies have started to report other patterns, such as correlations between species and genetic $\beta$-diversity (Papadopoulou et al. 2011; Baselga et al. 2013). Theoretical analyses, along the line followed here, are certainly required to evaluate their inferential power and to incorporate them in a spatially explicit neutral theory of SGDCs.

F.L. is supported by a fellowship from AgroParistech. This work was supported by the Affairs project of the Agence Nationale de la Recherche (BioAdapt program; to P.D.). Thanks are due to CNRS for supporting the work of P.J., P.D., and F.M.

Acknowledgments

Many thanks to U. Berger, J. Bronstein, and three anonymous referees for their useful comments on the manuscript. We thank the Centre d’Ecologie Fonctionnelle et Evolutive and M.-C. Quidoz (Système d’Information pour l’Ecologie platform) for access to computational resources.

Literature Cited


Appendix A from F. Laroche et al., “A Neutral Theory for Interpreting Correlations between Species and Genetic Diversity in Communities” (Am. Nat., vol. 185, no. 1, p. 59)

Mathematical Analysis of the Neutral Model of SGDC

Timescale Assumptions about Speciation and Mutation in the Neutral Model of the SGDC

We first define the local timescale in the model. We then provide conditions on model parameters that allow us to assume constant relative abundances of species in the regional pool at the local timescale as well as those under which mutation can be neglected.

Definition and Duration of Local Timescale

The whole theoretical treatment of the model is done considering the associated coalescent process. In other words, we focus on lineages of a given sample of individuals going from the present toward the past. Backward in time, lineages within local sites can either merge (a coalescence event) or go back in the regional pool (an immigration event). The number of lineages within sites decreases because of coalescence and immigration events until reaching zero. The event that leads to this state is necessarily an immigration event, because the last lineage in the site has no other lineage to merge with. We refer to this event as “first immigration.” We define the local timescale as the time span that goes from the first immigration to the present.

In a local site with parameters \( K \) and \( m \), a lineage has a probability of undergoing a death-birth event and \( m/K \) reaching immigration at each time step. Then, assuming that \( S \) individuals have been sampled at the present time, one gets a rough upper bound of local timescale \( T_{\text{loc}} \)

\[
E[T_{\text{loc}}] \leq \frac{SK}{m} \delta_{\text{loc}} = \frac{S}{m} \Lambda ,
\]

(A1)

where \( \delta_{\text{loc}} \) is the expected duration of a time step in the site and \( \Lambda \) is the life expectancy of an individual. In the following, we assume that \( S \) is not very high (\( S = 50 \)).

Condition for Getting a Constant Composition of the Regional Species Pool at the Local Timescale

Making the assumptions that (i) the regional pool of individuals undergoes a death-replacement dynamic, (ii) \( K_{\text{reg}} \gg 1 \), and (iii) speciation can be neglected ensures that the stochastic dynamics of species \( i \) relative abundance \( (f_i) \) verifies

\[
\begin{align*}
P(f_i(t+1) = f_i(t) + df | f_i(t)) &= (1 - f_i(t))f_i(t) \\
P(f_i(t+1) = f_i(t) | f_i(t)) &= (1 - f_i(t))^2 + f_i(t)^2 \\
P(f_i(t+1) = f_i(t) - df | f_i(t)) &= f_i(t)(1 - f_i(t))
\end{align*}
\]

(A2)

where \( df = 1/K_{\text{reg}} \) and time is measured in regional death replacement cycles (i.e., on the scale of \( \Lambda/K_{\text{reg}} \)).

Equation (A2) yields \( f_i \), conditional variance through time

\[
\forall[f_i(t+z)|f_i(t)] = (1 - f_i(t))f_i(t) \left( 1 - 2 \frac{K_{\text{reg}}}{K_{\text{reg}} - 1} \left( df \right)^2 \right)^{K_{\text{reg}}/(z/\Lambda)} ,
\]

(A3)

where \( z \) is a time span measured in real time.
Equation (A3) yields a sufficient condition to consider that species $i$ relative abundance is constant over time span $z$

\[1 - \left[1 - 2\left(\frac{K_{reg}}{K_{reg} - 1}\right)^{K_{reg}/z(A)}\right] \approx 1 \Rightarrow \left[1 - 2\left(\frac{K_{reg}}{K_{reg} - 1}\right)^{K_{reg}/z(A)}\right] \approx 1 \Rightarrow e^{-2z/(K_{reg} - 1)} \approx 1 \Rightarrow \frac{2z}{\Lambda K_{reg}} \ll 1.\]  

(A4)

Coming back to the model and combining equation (A1) and condition (A4), one obtains a sufficient condition on local immigration rate to consider the structure of the regional species pool as a constant at the local timescale

\[1 \ll K_{reg} m.\]  

(A5)

A sufficient condition justifying that the impact of speciation on regional relative abundances can be neglected at the local timescale is

\[\frac{S}{m} \frac{\Lambda K_{reg} p}{K_{reg}} \ll K_{reg} \Rightarrow p \ll m,\]  

(A6)

where $p$ would be a punctual “per cycle” speciation rate. We assume that both conditions are verified in this work.

**Condition to Neglect Mutation at Local Timescale**

We consider an infinite-alleles model with mutation rate $\mu$. The probability of observing no mutation at the local scale during coalescence is less than $(1 - \mu)^{5m} \approx e^{-5m \mu}$. The condition $m \gg \mu$ ensures that mutation can be locally neglected. We refer to this situation as the “weak mutation regime.” By analogy with condition (A6) above, this condition also ensures that the allelic frequencies in the regional population of the focal species can be considered constant at the local timescale.

We also consider the “strong mutation” situation, in which $m \approx \mu$. Assuming that the focal species is not rare (a necessary assumption to perform sampling, as shown in app. A), inequality (A1) implies $\theta = K_{reg} f_{i} \mu \gg 1$. The regional population of the focal species shows enough genetic variability to ensure that any new immigrant in local sites harbors an allele that never occurred locally before.

**Complete Analysis of the Neutral Model of the SGDC under the Weak Mutation Regime**

We consider local sites with parameters ($K, m$), and we define $I = (K - 1)m/(1 - m)$. We assume that condition (A6) is verified. First, we establish the joint sampling formula for species and genetic samples and show that both sample composition are probabilistically independent given $s_{e}$. Then, we develop the case of species richness and genetic richness. We finally discuss how these theoretical results may be biased when neglecting the assumption that $s_{e} \geq k$ in the site.

**Joint Sampling Formula and Independence of $s$ and $u$**

When the condition $s_{e} \geq k$ is neglected, species sample structure follows the distribution described in Etienne and Alonso (2005),

\[P(s) = \frac{S!}{\prod_{i=1}^{s} s_{i}!} \times \frac{\prod_{i=1}^{u} (I_{i}/f_{i})_{s_{i}}}{(I)_{s_{e}}},\]  

(A7)
where \( x_{o(i)} = \prod_{i=0}^{\infty} (x + i) \). In the following, we note \( \mathbb{P}_s \) probabilities conditionally on \( s \). Using the sampling formula of Etienne and Alonso (2005), we derive the probability of \( t \) composition conditionally on \( s \),

\[
\mathbb{P}_{s, t}(t) = \frac{\mathbb{P}(s \cap t)}{\mathbb{P}(s)} = \frac{s!(I(1 - f_s)_s)_{s-1} \prod_{i=1}^{s} (I f_e g_e) / (s - s_i) \prod_{i=1}^{s} t_i! / (I)_s}{s!(I(1 - f_s)_s)_{s-1} / (s - s_i) \prod_{i=1}^{s} t_i! / (I)_s} \tag{A8}
\]

where \( g = (g_1, g_2, \ldots) \) is the allelic frequencies marker of the focal species (and \( g_1 > g_2 > \ldots g_s \)). It follows a Ewens distribution as described in Ewens and Tavare (2006). This distribution is, in principle, infinite, but we truncated here at rank \( A \) (very large), leading to a renormalized frequency distribution \( g = (g_1, g_2, \ldots g_s) \). This allows us to use the sampling formulae of Etienne and Alonso (2005). Equation (A8) shows that, when \( s \) is known, the composition of \( t \) follows a sampling formula that is similar to the one described in Etienne and Alonso (2005). This implies that \( t \) structure verifies the “subsampling property” so that \( u \) composition can be inferred as follows:

\[
\mathbb{P}_u(u) = \frac{k!}{\prod_{i=1}^{k} u_i!} \times \frac{\prod_{i=1}^{k} (I f_e g_e) / (I)_s}{(I(1 - f_s)_s)_{s-1} / (s - s_i) \prod_{i=1}^{s} t_i! / (I)_s} \tag{A9}
\]

Equation (A9) does not depend on \( s \), so the unconditional distribution of \( u \) is given by

\[
\mathbb{P}(u) = \frac{k!}{\prod_{i=1}^{k} u_i!} \times \frac{\prod_{i=1}^{k} (I f_e g_e) / (I)_s}{(I(1 - f_s)_s)_{s-1} / (s - s_i) \prod_{i=1}^{s} t_i! / (I)_s} \tag{A10}
\]

In particular, equation (A10) shows that the relative strength of migration compared with drift in the local populations of the focal species is \( I f_e \) instead of \( I \) for the whole community.

Using the sampling formula from Etienne and Alonso (2005), one can verify that

\[
\mathbb{P}_s(s \cap t) = \frac{S! / \prod_{i=1}^{s} s_i! \prod_{i=1}^{s} t_i! \times \prod_{i=1}^{s} (I f_e g_e) / (I)_s / (I(1 - f_s)_s)_{s-1} / (s - s_i) \prod_{i=1}^{s} t_i! / (I)_s}{S! / \prod_{i=1}^{s} s_i! / (s - s_i) ! \prod_{i=1}^{s} t_i! / (I)_s} = \frac{(S - s_i)!}{\prod_{i=1}^{s} s_i!} \times \frac{\prod_{i=1}^{s} (I f_e g_e) / (I)_s}{(I(1 - f_s)_s)_{s-1} / (s - s_i) \prod_{i=1}^{s} t_i! / (I)_s} \tag{A11}
\]

Equation (A11) shows that the composition of the remaining part of the species sample \( s \) is independent from \( t \) composition given \( s \). With \( u \) being a subsample of \( t \), this result also applies to \( u \)

\[
\mathbb{P}_u(s \cap u) = \mathbb{P}_u(s) \mathbb{P}_u(u). \tag{A12}
\]

As shown in equation (A9), \( \mathbb{P}_u(u) \) does not depend on \( s \). Thus, taking the expectation of equation (A12), we obtain the independence of \( s \) and \( u \) compositions

\[
\mathbb{P}(s \cap u) = \mathbb{P}(s) \mathbb{P}(u). \tag{A13}
\]

Species Richness and Allelic Richness

Using the sampling formula of Etienne and Alonso (2005), the probability \( \mathbb{P}_e(s) \) of observing a composition of the species sample given the number of individuals belonging to species \( e \) can be written as

\[
\mathbb{P}_e(s) = \frac{(S - s_i)!}{\prod_{i=1}^{s} s_i!} \times \frac{\prod_{i=1}^{s} (I f_e g_e) / (I)_s}{(I(1 - f_s)_s)_{s-1} / (s - s_i)}. \tag{A14}
\]
From equation (A14), we derive the probability $p_i$ of species $i$ occurring in the sample, given the number of individuals belonging to species $e$

$$p_i = 1 - \frac{[I(1 - f_e - f_i)]_{[s-e]}}{[I(1 - f_e)]_{[s-e]}}.$$ (A15)

Equation (A15) implies that the expectation for species richness in the species sample $\mathbf{s}$, given the number of individuals belonging to species $e$, is equal to

$$\overline{R}_{\text{spec}}(s, I) = 1 + \sum_{i=1}^{h} p_i = B - \sum_{i=1}^{h} \frac{[I(1 - f_i)]_{[s-i]}}{[I(1 - f_e)]_{[s-e]}}.$$ (A16)

It can be shown on the basis of a simple development of equation (A16) that $\overline{R}_{\text{spec}}(s, I)$ is a decreasing function of $s_e$.

When not specifying the value of $s_e$ and neglecting the assumption that $s_e \geq k$, the species sample structure follows the distribution described in equation (A7). Therefore, the probability of species $i$ occurring in the sample $q_i$ is equal to

$$q_i = 1 - \frac{[I(1 - f_i)]_{[\phi]}}{[I(\phi)]_{[\phi]}}.$$ (A17)

Equation (A17) implies that the expectation for species richness in the species sample $\mathbf{s}$ is equal to

$$\overline{R}_{\text{spec}}(I) = \sum_{i=1}^{h} q_i = B - \sum_{i=1}^{h} \frac{[I(1 - f_i)]_{[\phi]}}{[I(\phi)]_{[\phi]}}.$$ (A18)

Equation (A18) is the expectation on species richness when no condition is put on the number of individuals belonging to species $e$. Little calculation from equation (A18) shows that $\overline{R}_{\text{spec}}(I)$ is an increasing function of $I$.

Equation (A10) shows that $u$ sampling formula can be obtained from $s$ sampling formula by replacing $I$ by $I_{e}$, $S$ by $k$, and $f$ by $g$. Therefore, expected allelic richness in a site can be directly deduced from expected species richness in a site given by equation (A18),

$$\overline{R}_{\text{all}}(I) = A - \sum_{i=1}^{h} \frac{[I_{f}(1 - g_i)]_{[\phi]}}{[I_{f}(\phi)]_{[\phi]}},$$ (A19)

and $\overline{R}_{\text{all}}(I)$ is an increasing function of $I$.

Effects of Neglecting the Assumption That $s_e \geq k$

Neglecting the assumption that $s_e \geq k$ leads to use equation (A7) instead of the true sampling formula that writes

$$P_I(\mathbf{s}) = \begin{cases} 0 & \text{if } s_e < k \\ \frac{1}{1 - P} \times \frac{S!}{\prod_{i=1}^{h} s_i!} \times \frac{\prod_{i=1}^{h} [I_{f}(1 - g_i)]_{[\phi]}}{[I(\phi)]_{[\phi]}} & \text{if } s_e \geq k \end{cases},$$ (A20)

where $P = \sum_{j=0}^{k-1} P_j$ is the probability of observing fewer than $k$ individuals in the species sample, and $P_j$ is the probability of observing exactly $j$ individuals of species $e$ in the sample.

We chose to work with equation (A7), because it leads to a more tractable derivation, but one needs to quantify the potential bias of such an approximation. The condition $s_e \geq k$ has little effect on species sample when $P \ll 1$. To assess when this occurs, we place ourselves under the diffusion limit, where $K \to +\infty; m \to 0; Km \to I$ finite.

Assuming that $S \gg 1$ and defining $\chi$ as the proportion of the species sample occupied by species $e$ (i.e., $s_e/S$) density $\phi$ of $\chi$ writes (Wright 1931)

$$\phi(\chi; I_{f_e}) = \begin{cases} \frac{\Gamma(I)}{\Gamma(g_e)\Gamma(I(1 - f_e))} \chi^{g_e-1}(1 - \chi)^{R(1 - f_e) - 1} & \text{if } \chi \in [0, 1] \\ 0 & \text{else} \end{cases},$$ (A21)

and probability $P$ defined above asymptotically verifies

$$P = \int_0^{(g_e)} \phi(\chi; I_{f_e}) d\chi.$$ (A22)
Appendix A from F. Laroche et al., Neutral Theory across Diversity Levels

\( P \) increases with \( k \). According to the properties of the \( \beta \) distribution, one can predict that \( P \) decreases with \( f_e \).

For given values of \( k/S \) and \( f_e \), figure A1 shows that there exists a threshold of \( I \) above which \( P \) is negligible (i.e., below .05). It also shows that, for a given value of \( k/S \), this threshold is a decreasing function of \( f_e \). In the whole study, we consider situations in which \( P \) is negligible. To perform fast simulations, we focus on the case where \( k = 5 \) and \( S = 50 \) (i.e., a \( k/S \) ratio of 0.1; fig. A1A). We consider a focal species with regional relative abundance of 0.8 so that we have a threshold for \( I \) of 1. With those parameters, \( s \) composition could be considered to follow the distribution described in equation (A7).

Analysis of the Variation in \( \bar{R}_{all}(I, m) \) under the Strong Mutation Regime

Within local sites, \( R_{all} \) depends on two features: the number of migrants that funded the genetic sample \( u \), called \( N_u \) (for number of ancestors), and the number of mutation events that occur in the coalescence tree associated to \( u \) (that we call the \( u \) tree), called \( M \). \( R_{all} \) is an increasing function of \( N_u \) (\( M \) being kept constant) and an increasing function of \( M \) (\( N_u \) being kept constant).

\( N_u \) only depends on the topology of the \( u \) tree (i.e., which lineages coalesce and which ones emigrate), which is exclusively driven by parameter \( I \). Expectation of \( N_u \) is given by

\[
N_u = \sum_{n=1}^{s} \frac{I f_e}{I f_e + n - 1}.
\]  

The expected number of birth-death events in a lineage of the \( s \) tree during the phase where the \( s \) tree has exactly \( n \) local lineages \( \bar{B}_n \) is \( \bar{B}_n = [1/(nm)][I/(I + n - 1)] \). Thus, the overall expected number of mutation events in the lineage during the same phase is given as \( \bar{F}_n = [\mu/(nm)][I/(I + n - 1)] \). The \( u \) tree is a subset of the branches of the \( s \) tree. We define the set of random variables \( D = \{D_1, \ldots, D_s\} \), where \( D_i \) is the number of branches of the \( u \) tree during the phase where the \( s \) tree has \( i \) lineages. The expectation of \( M \) writes

\[
M = \sum_{n=1}^{s} \bar{D}_n \bar{F}_n = \frac{\mu}{m} \sum_{n=1}^{s} \bar{D}_n \left[ \frac{I}{I + n - 1} \right],
\]  

where \( \bar{D}_n \) is the expectation of \( D_n \) and depends on \( s \) tree topology (i.e., on \( I \) parameter) and \( f_e \).

We assume that \( R_{all} \) is an increasing function of \( N_u \) and \( M \). Then, equations (A23) and (A24) imply that, \( I \) being kept constant, \( R_{all} \) is a decreasing function of \( m \) because \( \bar{M} \), the mean number of mutation events in the \( u \) tree, is proportional to the mutation to migration ratio \( \mu/m \), whereas the number of ancestors \( N_u \) is unaffected by the value of \( m \). Formally, we obtain

\[
\frac{\partial R_{all}(I, m)}{\partial m} > 0.
\]  

Under the weak mutation regime, similarly to species level, we have \( \frac{\partial R_{all}(I)\partial I}{\partial m} > 0 \). This pattern may change when mutation is strong because of the impact of \( I \) on \( \bar{D}_n \) values. However, on the basis of a continuity argument, we expect this property to be maintained until a certain level of mutation rate, so that \( \frac{\partial R_{all}(I, m)\partial I}{\partial m} > 0 \). Unfortunately, we could not establish the generality of this result.
Figure A1: Value of $P$ as a function of $I$ and $f_e$ for various values of $k/S$: (A) $k/S = 0.1$, (B) $k/S = 0.01$, and (C) $k/S = 0.001$. Bold line materializes the threshold of $I$ above which $P < .05$ as a function of $f_e$. 
Appendix B from F. Laroche et al., “A Neutral Theory for Interpreting Correlations between Species and Genetic Diversity in Communities” (Am. Nat., vol. 185, no. 1, p. 59)

Modeling the Landscape through a “Discretized” Bivariate Gaussian Distribution of $K$ and $m$

Building $(K, m)$ Distribution

We model $(K, m)$ distribution across sites as follows:

$$\log (K - 1), \log \left( \frac{m}{1 - m} \right) = [f(X), Y], \quad (B1)$$

where $(X, Y)$ is a Gaussian vector with mean $(\alpha_K, \alpha_m)$ and variance/covariance matrix

$$\Sigma = \begin{pmatrix} \sigma_K^2 & \rho \sigma_K \sigma_m \\ \rho \sigma_K \sigma_m & \sigma_m^2 \end{pmatrix},$$

$\rho \in [-1, 1]$, and $f(x) = \log \left( \left\lfloor 10^x \right\rfloor + 1 \right)$, where $\left\lfloor \cdot \right\rfloor$ is the integer part operator. Here we worked with given $\sigma_K^2$ and explored variation in $\sigma_m^2$ and $\rho_{km}$.

Linking $(K, m)$ and $(I, m)$ Distribution

$\sigma_I^2 = \text{Var} \{ \log (I) \}$ and $C_{im} = \text{Cov} \{ \log (I), \log [m/(1 - m)] \}$ can be linked to $(K, m)$ distribution. Considering that

$$\log (I) = \log (K - 1) + \log \left( \frac{m}{1 - m} \right), \quad (B2)$$

and assuming that discretization through $f$ has no important effect, which is likely to be true when $K$ is high enough (e.g., when $\mu_K$ is higher than 2 and $\sigma_K < \mu_K/2$), $\sigma_I^2$ and $C_{im}$ write

$$\sigma_I^2 = \sigma_K^2 + \sigma_m^2 + 2 \rho_{km} \sigma_K \sigma_m, \quad (B3)$$

$$C_{im} = \rho_{km} \sigma_K \sigma_m + \sigma_m^2, \quad (B4)$$

Equation (B3) shows that a high variance in $K$ and $m$ does not necessarily lead to a high variance in $I$ if $\rho_{km}$ is strongly negative. For a given value of $\sigma_I^2$, equations (B3) and (B4) yield

$$C_{im} = \frac{\sigma_I^2 + \sigma_m^2 - \sigma_K^2}{2}, \quad (B5)$$

Equation (B5) implies that, for a given value of $\sigma_I^2$ and $\sigma_K^2$, landscapes with higher variance in $m$ harbor stronger $C_{im}$, as illustrated in figure B1, whereas for a given value of $\sigma_I^2$ and $\sigma_m^2$, landscapes with lower variance in $K$ harbor stronger $C_{im}$.

For a given value of $\sigma_K^2$, increasing $\sigma_m^2$ keeping $\sigma_I^2$ constant can be obtained by compensatory variation of $\rho_{km}$ verifying

$$\frac{\partial \sigma_I^2}{\partial \sigma_m} = 2 \sigma_m + 2 \rho_{km} \sigma_K + 2 \frac{\partial \rho_{km}}{\partial \sigma_m} \sigma_K \sigma_m$$

$$0 \Leftrightarrow \frac{\partial \rho_{km}}{\partial \sigma_m} = \frac{C_{im} - \sigma_m + \rho_{km} \sigma_K}{\sigma_K \sigma_m}, \quad (B6)$$

Equation (B6) implies that, among landscapes sharing the same values of $\sigma_I^2$ and $\sigma_K^2$ and harboring positive $C_{im}$ values, $\sigma_m^2$
and $\rho_K$ are negatively related, and a higher value of $\sigma^2_a$ along with a lower value of $\rho_K$ leads to a higher $C_{lm}$ value (fig. B1). By contrast, among landscapes sharing the same values of $\sigma^2_I$ and $\sigma^2_K$ and harboring negative $C_{lm}$ value, $\sigma^2_m$ and $\rho_K$ are positively related, and higher values of $\sigma^2_m$ and $\rho_K$ lead to higher $C_{lm}$ value.

![Figure B1](image.png)

**Figure B1:** Features of various distributions of $\{\log(K-1), \log(m/(1-m))\}$ with the same value of $\sigma^2_I$. Distributions were generated by exploring values of $\sigma^2_a$ while keeping $\alpha_K$, $\alpha_m$, and $\sigma^2_K$ constant, $\rho_K$ was adjusted so as to keep $\sigma^2_I$ constant in every distribution, using equation (B3). In A, distributions are represented as dots in the $(\sigma^2_m, \rho_K)$ plan. The levels of gray indicate the corresponding value of $C_{lm}$. In B, the same distributions were represented using ellipsoids encompassing a 0.95 probability zone around the distribution mode. The levels of gray indicate the corresponding value of $C_{lm}$ according to the color key described in A. Constant parameters were set to $\alpha_K = 3$, $\alpha_m = -3$, $\sigma^2_K = 3$, and $\sigma^2_I = 0.5$. 

Appendix C from F. Laroche et al., “A Neutral Theory for Interpreting Correlations between Species and Genetic Diversity in Communities” (Am. Nat., vol. 185, no. 1, p. 59)

Decomposition of Covariance

Let \( X, Y, Z = (Z_1, Z_2, \ldots, Z_s) \) be random variables. We use the following lemma:

\[
\text{Cov}(X, Y) = \text{Cov}(\bar{X}(Z), \bar{Y}(Z)) + \mathbb{E}[\text{Cov}_Z(X, Y)],
\]

(C1)

where \( \bar{X}(Z) \) is the conditional expectation of \( X \) knowing \( Z \) values, \( \bar{Y}(Z) \) is the conditional expectation of \( Y \) knowing \( Z \) values, and \( \text{Cov}_Z(X, Y) \) is the conditional covariance between \( X \) and \( Y \) knowing \( Z \) values. \( \bar{X}(Z), \bar{Y}(Z), \) and \( \text{Cov}_Z(X, Y) \) are then three deterministic functions of \( Z \) values.

Applying equation (C1) with \( X = R_{\text{spec}}, Y = R_{\text{all}}, Z = (K, m) \) yields equation (1) of the main text,

\[
\text{Cov}(R_{\text{spec}}, R_{\text{all}}) = \text{Cov} \left[ \bar{R}_{\text{spec}}(K, m), \bar{R}_{\text{all}}(K, m) \right] + \mathbb{E}[\text{Cov}_{K, m}(R_{\text{spec}}, R_{\text{all}})].
\]

Equation (C1) can be applied conditionally to the fact that some other variables \( U = (U_1, \ldots, U_l) \) are known, such that

\[
\text{Cov}_U(X, Y) = \text{Cov}_U[\bar{X}(Z, U), \bar{Y}(Z, U)] + \mathbb{E}[\text{Cov}_{Z|U}(X, Y)](U).
\]

(C2)

In equation (C2), \( \text{Cov}_U \) means conditional covariance knowing \( U \) values and \( \bar{X}(Z, U) \) and \( \bar{Y}(Z, U) \) are the conditional expectation of \( X \) and \( Y \) knowing both \( Z \) values and \( U \) values. As \( U \) is supposed to be known, we consider \( \bar{X}(Z, U) \) and \( \bar{Y}(Z, U) \) as deterministic functions of random variables contained in \( Z \). \( \text{Cov}_U(\bar{X}(Z, U), \bar{Y}(Z, U)) \) is then the covariance between these functions for a given \( U \) when \( Z \) varies.

\( \text{Cov}_{U,Z}(X, Y) \) is the covariance of \( X \) and \( Y \) knowing \( U \) and \( Z \) values. As \( U \) is known, we consider \( \text{Cov}_{U,Z}(X, Y) \) as a deterministic function of random variables contained in \( Z \). \( \text{Cov}_{U,Z}(X, Y)(U) \) is then the expectation for a given \( U \) of this function, taken over all possible values of \( Z \).

Applying equation (C2) with \( X = R_{\text{spec}}, Y = R_{\text{all}}, Z = s_r, \) and \( U = (K, m) \) yields

\[
\text{Cov}_{K, m}(R_{\text{spec}}, R_{\text{all}}) = \text{Cov}_{K, m} \left[ \bar{R}_{\text{spec}}(s_r, K, m), \bar{R}_{\text{all}}(s_r, K, m) \right] + \mathbb{E}[\text{Cov}_{K, m,s_r}(R_{\text{spec}}, R_{\text{all}})|(K, m)].
\]

(C3)

We show in appendix A that \( R_{\text{spec}} \) and \( R_{\text{all}} \) are probabilistically independent when \( K, m, \) and \( s_r \) are known, which implies for any value of \( K, m, \) and \( s_r \),

\[
\text{Cov}_{K, m,s_r}(R_{\text{spec}}, R_{\text{all}}) = 0.
\]

(C4)

Applying equation (C3), equation (C4) simplifies to

\[
\text{Cov}_{K, m}(R_{\text{spec}}, R_{\text{all}}) = \text{Cov}_{K, m} \left[ \bar{R}_{\text{spec}}(s_r, K, m), \bar{R}_{\text{all}}(s_r, K, m) \right].
\]

(C5)

Under the weak mutation regime, \( \bar{R}_{\text{all}} \) does not depend on \( s_r \) (app. A). Then equation (C5) yields

\[
\text{Cov}_{K, m}(R_{\text{spec}}, R_{\text{all}}) = 0.
\]

(C6)

Under the strong mutation regime, \( \bar{R}_{\text{spec}}(s_r, K, m) \) decreases with \( s_r \), whereas \( \bar{R}_{\text{all}}(s_r, K, m) \) increases with \( s_r \). Thus equation (C5) yields

\[
\text{Cov}_{K, m}(R_{\text{spec}}, R_{\text{all}}) < 0.
\]

(C7)
Author-supplied PDFs: Pseudo-code of the coalescence algorithm in a site

We consider the genealogy of the species sample $s$ backward in time. The algorithm is initialized considering present individuals in separate lineages, with unknown species and allele labels. Then lineages merge (coalescence event) or leave the local site toward the regional pool (migration event) as time goes back. Assuming stability of species composition in the regional community and drift/mutation equilibrium in the regional population of the focal species allows determining both species and allele of the first migrant of each lineage, called the ancestor. These species and genetic labels can then be propagated to all individuals of the present that belong to this lineage, accounting for local mutation events under the strong mutation assumption.

*Algorithm under weak mutation*

This algorithm builds from the coalescence algorithm proposed by (Etienne 2005). In particular, alleles of ancestors (see part 3) can be determined by a sequential urn algorithm that was described in (Ewens and Tavaré 2004).

```plaintext
//0 - Initializing

K, m, θ;

l := (K-1)*m/(1-m);

ANCESTRAL_LINEAGES := NULL;

FOCAL_ANCESTRAL_LINEAGES := NULL;

LINEAGES := {{Individual 1}, {Individual 2},..., {Individual S}};

FOR (every individual IND in LINEAGES){
```
IND.speciesLabel := NULL;

IND.allelicLabel := NULL;
}

NB_LIN:= S;

NEW_ALLELE:=0;

//1 – Determining groups of individuals that share the same immigrant ancestor

WHILE (NB_LIN>=2) {

  Choose at random an element in LINEAGES, called L1;

  Remove L1 from LINEAGES

  Define X1, a Bernoulli random variable with probability I/(I+NB_LIN-1);

  IF(X1==1)

  THEN add L1 to ANCESTRAL_LINEAGES

  ELSE{

  Choose at random an element in LINEAGES, called L2;

  Remove L2 from LINEAGES;

  Define L3, the lineage resulting from merging individuals of L1 and L2;

  Add L3 to LINEAGES;

  }

  Decrease NB_LIN by 1;
Add the last element of LINEAGES to ANCESTRAL_LINEAGES;

//2 – Assigning species to those groups

FOR (every element A of ANCESTRAL_LINEAGES) {

    Define SPE, a species label randomly chosen according to \( f \) distribution;

    FOR (every element IND of A) {
        IND.speciesLabel := SPE;
    }

    IF (SPE is the label of focal species) {
        THEN add A to FOCAL_ANCESTRAL_LINEAGES;
        ELSE{}
    }
}

//3 - Genotyping individuals belonging to the focal species

FOR (INDEX going from 1 to the length of FOCAL_ANCESTRAL_LINEAGES) {

    Define FA, the INDEX\(^{th}\) element of FOCAL_ANCESTRAL_LINEAGES;

    Define IS_NEWALL, the realization a Bernoulli random variable with probability \( \frac{\Theta}{\Theta + INDEX-1} \)

    IF (IS_NEWALL == 1) {
        THEN {
            
        }
    }
FOR(every element IND of FA){

    IND.allelicLabel := NEW_ALLELE;

}

Increase NEW_ALLELE by 1;

}

ELSE{

    Define FA2, a lineage randomly chosen among the 1 to (INDEX-1)th elements of FOCAL_ANCESTRAL_LINEAGES;

    Define ALLELE, the allelic label of the elements of FA2;

    FOR(every element IND of FA){

        IND.allelicLabel := ALLELE;

    }

}

}
some individuals belonging to non-focal species may harbor an allele. These allelic states are not taken into account in our analysis.

//0 - Initializing

K, m, θ, μ;

l := (K-1)*m/(1-m);

FOR (i going from 1 to S) {
    (Individual_i).speciesLabel := NULL;
    (Individual_i).allelicLabel := NULL;
}

LINEAGES := {{Individual_1}, {Individual_2},..., {Individual_S}};

NB_LIN := S;

ANCESTRAL_LINEAGES := NULL;

FOCAL_ANCESTRAL_LINEAGES := NULL;

NEW_ALLELE := 0;

//1 – Determining groups of individuals that share the same immigrant ancestor

WHILE (NB_LIN >= 2) {
    Define T, the realization of a geometric random variable with parameter m+(1-m)*(NB_LIN-1)/(K-1)

    Define NB_MUT, the realization of a Binomial random variable with parameters (μ, T);
FOR(i going from 1 to NB_MUT){
    Define L1, a randomly chosen lineage of LINEAGES
    FOR(every element IND of L1){
        IF(IND.allelicLabel is NULL)
            THEN IND.allelicLabel:=NEWALLELE
        ELSE{}
    }
    Increase NEW_ALLELE by 1;
}

Define L2, an randomly chosen element in LINEAGES;
Remove L2 from LINEAGES

Define $X_1$, the realization of a Bernoulli random variable with probability $I/(I+NB\_LIN-1)$;

IF($X_1$==1)
    THEN add L2 to ANCESTRAL\_LINEAGES
ELSE{
    Choose at random an element in LINEAGES, called L3;
    Remove L3 from LINEAGES;
Define L4, the lineage resulting from merging individuals of L2 and L3;

Add L4 to LINEAGES;

}

Decrease NB_LIN by 1;

}

Add the last element of LINEAGES to ANCESTRAL_LINEAGES;

//2 – Assigning species to those groups

FOR (all the elements of ANCESTRAL_LINEAGES){

    Define A, the current element of ANCESTRAL_LINEAGES;

    Define SPE, a species label randomly chosen according to \( f \) distribution;

    Assign SPE to all the elements of A;

    IF (SPE is the label of focal species)

        THEN add A to FOCAL_ANCESTRAL_LINEAGES;

    ELSE{}

}

//3 - Genotyping individuals belonging to focal species

FOR (every lineage FA in FOCAL_ANCESTRAL_LINEAGES){

    FOR(every element IND of FA){

        IF(IND.allelicLabel is NULL)
THEN IND.allellicLabel:=NEWALLELE

ELSE{}

IF (at least 1 individual of FA had a NULL allelic label)

THEN increase NEW_ALLELE by 1;

ELSE{}

}