CHAPTER EIGHT

# The importance of phylogenetic structure in biodiversity studies

JÉRÔME CHAVE
*Université Paul Sabatier, Toulouse*
GUILLEM CHUST
*Université Paul Sabatier, Toulouse*
CHRISTOPHE THÉBAUD
*Université Paul Sabatier, Toulouse*

## Introduction

A central goal of biodiversity research is to understand processes of species coexistence at different spatial and temporal scales. Much empirical research has revolved around documenting patterns of species abundance and distribution with sound sampling techniques and statistics (May, 1975; Magurran, 1988; Krebs, 1998). Such data are of tremendous importance not only for documenting current biological diversity patterns, but also for testing fundamental ecological theories (Ricklefs & Schluter, 1993; Brown, 1995; Hubbell, 2001). A common feature of these approaches is the emphasis placed on species as the appropriate currency for quantifying biological diversity. However, documenting species diversity often represents a considerable practical challenge. First, no one exactly knows the total number of extant species on Earth (Erwin, 1982; May, 1994; Novotný *et al.*, 2002; Alroy, 2002). Second, in any given sample, a sizeable fraction of the individuals may represent previously undescribed species, as is especially the case for lesser known groups, like plants in tropical forests, insects or protists. Third, species recognition usually relies upon a set of morphological cues which are not always observable. Thus, many individuals within a sample cannot be reliably assigned to previously described species. This is obvious in microbial communities, where different operational taxonomic units (OTUs) can only be distinguished by DNA screening or other molecular methods (e.g. see Suau *et al.*, 1999 for a study of the microbial diversity of the human gut, and Green & Bohannan, this volume). This issue is also serious in macroscopic organisms. For example, in spite of enormous efforts by botanists, as many as 40% of South American herbarium specimens may be improperly identified (M. J. Hopkins, personal communication). In general, there is considerable uncertainty concerning the quality of diversity surveys, and a consistent definition of OTUs is lacking.

These technical difficulties hamper reliable estimation of the local species diversity in ecological communities, and regional maps of local species diversity are likely to suffer from these limitations. This clearly is a limitation in conservation biology when ecosystems are valued proportionally to their local species richness (Faith, 1994). Assessing landscape-scale diversity – or beta diversity (*sensu* Whittaker, 1972) – is even more difficult than alpha diversity. Beta diversity measures the turnover in species composition – or species overlap – among sites and within a landscape (Gaston *et al.*, this volume). Species overlap is meaningful only if the taxa are consistently delimited across the landscape. To circumvent such problems, some tropical forest plant researchers have limited their investigations to better-known taxonomic groups, such as palms (Kahn & Meija, 1991; Clark *et al.*, 1995; Vormisto *et al.*, 2004), ferns and fern allies (Tuomisto & Poulsen, 2000; Tuomisto *et al.*, 2003), in the hope that these groups are good indicators of diversity.

Another approach to this problem is to consider that the diversity represented by species richness and species overlap is only part of the overall evolutionary diversity in communities. Species are of uneven importance in view of their place in the tree of life, if only because they may represent variable amounts of evolutionary history (Vane-Wright *et al.*, 1991; Faith, 1994; Nee & May, 1997). One simple way of measuring this taxonomic diversity is to compute species-to-genera ratios: if the ratio is high, then on average every genus is represented by many species and the loss of any one species should not result in any important evolutionary loss. However, species-to-genera ratios may not appropriately summarize the phylogenetic information if the corresponding phylogenetic tree is unbalanced (Yule, 1925; Mooers & Heard, 1997; Webb & Pitman, 2002), and because they are dependent on sample size (Gotelli & Colwell, 2001). Extensions of this approach consist in making full use of phylogenetic tree topology in measures of biological diversity and of overlap. For instance, methods have been developed to assess the conservation value of species (Faith, 1994; Nee & May, 1997; Pavoine *et al.*, 2005). More recently, it has been realized that measures of phylogenetic diversity could be used to evaluate the evolutionary originality of a whole community (Webb, 2000; Webb *et al.*, 2002) and to test fundamental biodiversity theories. For instance if a community is made of species randomly drawn from a regional species pool – as should be the case if the neutral theory of biodiversity (Hubbell, 2001) holds – then they should evenly represent the clades of the regional phylogeny. By contrast, if limiting similarity is predicted because of competitive exclusion of closely allied species, then the species should be overdispersed on the phylogeny (Webb *et al.*, 2002; Webb & Pitman, 2002). Thus integrating the evolutionary dimension into biodiversity research goes beyond a simple refinement of diversity indices. It provides new tests for comparing ecological processes.

Here, we define a simple and consistent measure of phylogenetic diversity, based on existing population genetics theory, and which has several mathematical advantages. Then we explore how making use of information on evolutionary relationships among species may aid in understanding why biodiversity varies within and across localities. To illustrate the use of the measures, we use data on South American plant communities assembled by A. H. Gentry (1982, 1988) and phylogenetic information from the Angiosperm Phylogeny Group (henceforth APG) consortium. Specifically, we address the following questions: (1) Is the phylogenetic measure of diversity well predicted by nonphylogenetic measures based on species diversity? (2) Does the phylogenetic structure have an influence on biodiversity patterns in these communities? (3) What is the spatial structure of phylogenetic diversity, and what are the determinants in its variation? Thus, our overall goal is to examine to what extent phylogenetics is informative to central questions of community ecology, and the interplay between ecological versus evolutionary explanations of biodiversity patterns (Ricklefs, 2004). We do not address the biogeographical implications of our analysis (Gentry, 1988; Prance, 1994), a topic deferred to a forthcoming publication.

## Theory

Population genetics has been instrumental in developing statistics characterizing population substructure (Wright, 1931; Ewens, 1972; Watterson, 1974; Nei, 1987; Lande, 1996). We first summarize common measures of diversity in population genetics, then establish the correspondence with diversity measures in community ecology. Let us consider $K$ distinct populations, and $x_{ki}$, the relative abundance of the $i$th allele in population $k$ for one locus having $S$ alleles. Then classical population genetics theory (e.g. Nei, 1987) utilizes the following measure of diversity between populations $k$ and $l$:

$$\overline{D}_{kl} = 1 - \sum_{i=1}^{S} x_{ki}x_{li}. \tag{8.1}$$

The basic measure of local gene diversity is the heterozygosity, the probability that, in a randomly mating population, two individuals have different alleles at the same polymorphic locus, or $\overline{D}_{kk} = 1 - \sum (x_{ki})^2 = \sum_{i,j,i\neq j} x_{ki}x_{kj}$. Nei (1973) showed that a measure of diversity between two populations, excluding intrapopulation diversity, would be

$$\overline{H}_{kl} = \overline{D}_{kl} - \frac{\overline{D}_{kk} + \overline{D}_{ll}}{2} = \frac{1}{2}\sum_{i=1}^{S}(x_{ki} - x_{li})^2. \tag{8.2}$$

This quantity may be best thought of as proportional to the squared Euclidean distance between site $i$ and site $k$. The total diversity across all populations can then be defined as

$$\overline{D}_T = \langle \overline{D}_{kk} \rangle_k + \langle \overline{H}_{kl} \rangle_{k,l} = 1 - \sum_{i=1}^{S} \left( \langle x_{ki} \rangle_k \right)^2, \tag{8.3}$$

where in general $\langle Y_k \rangle_k = \dfrac{1}{K} \sum_{k=1}^{K} Y_k$ is the mean of variable $Y_k$ over $k$ ($\langle H_{kl} \rangle_{k,l}$ is the mean over both $k$ and $l$). $\overline{H}_{kl}$ is defined by Eq. (8.2).

Lande (1996) already pointed out that this formalism could be ported from population genetics into biodiversity studies. Specifically, he suggested that if $x_{ki}$ is now interpreted as being the relative abundance of taxon $i$ in habitat $k$, then Eqs. (8.1) to (8.3) define measures of local and spatial diversity. To follow the terminology of community ecology, $\overline{D}_{kk}$ is usually called the Simpson index, and is a measure of alpha diversity, and $\overline{H}_{kl}$ is a measure of beta diversity (Whittaker, 1972). At the community level Eq. (8.3) can be interpreted as an *additive* partition of diversity into a within-locality (alpha-diversity) component $\overline{D}_\alpha = \langle \overline{D}_{kk} \rangle_k$ and a between-locality (beta-diversity) component $\overline{D}_\beta = \langle \overline{H}_{kl} \rangle_{k,l}$.

Our goal is to extend this formalism to the study of phylogenetically structured communities. A simple generalization of formula (8.1) for genetically structured populations is

$$D_{kl} = \sum_{i=1}^{S} \sum_{j=1}^{S} t_{ij} x_{ki} x_{lj}. \tag{8.4}$$

In the present work, $t_{ij}$ denotes the divergence time between taxon $i$ and taxon $j$, such that the maximal divergence time is set to unity. In population genetic theory, $t_{ij}$ denotes the fraction of nucleotide substitutions between haplotypes $i$ and $j$ (Nei, 1987). Haplotype refers here to a set of closely linked alleles inherited as a unit. The above formulas (8.1–8.3) correspond to the limiting case where the divergence time is assumed to take the maximal value ($t_{ij} = 1$) for each $i \neq j$, which is also called a "star-shaped" phylogeny (every pair of distinct taxa is equidistant). To generalize the above formulas, we also define the matrix $P_{ij} = 1 - t_{ij}$, such that $P_{ij}$ for each $i \neq j$ and 1 otherwise. In this case, $D_{kl} = 1 - \sum_{i=1}^{S} \sum_{j=1}^{S} P_{ij} x_{ki} x_{lj}$. Note that Pavoine *et al.* (2005) recently called this quantity the "quadratic entropy", based on some earlier work in statistics.

In (8.4) and subsequent formulas, $D$ and $H$ now refer to measures of phylogenetic diversity taking into account the distribution of divergence times between species pairs. Throughout the text, nonphylogenetic measures of diversity are distinguished from phylogenetic measures by the bar over the symbol. We can write the companion of Eq. (8.2):

$$H_{kl} = D_{kl} - (D_{kk} + D_{ll})/2 = \frac{1}{2}\sum_{i=1}^{S}\sum_{j=1}^{S}(x_{ki} - x_{li})P_{ij}(x_{kj} - x_{lj}) \tag{8.5}$$

and of Eq. (8.3):

$$D_T = \langle D_{kk}\rangle_k + \langle H_{kl}\rangle_{k,l} = 1 - \sum_{i=1}^{S}\sum_{j=1}^{S}\langle x_{ki}\rangle_k P_{ij}\langle x_{kl}\rangle_k, \tag{8.6}$$

where $\omega_k$ is defined as in Eq. (8.3). At the community level this formula can still be interpreted as an additive partition of phylogenetic diversity into alpha and beta diversity $D_T = D_\alpha + D_\beta$.

In principle, the term $t_{ij}$ could still be interpreted as a measure of among-species divergence in DNA sequence data. This is most convenient in micro-biology, where the OTUs are directly defined by a DNA sequence (Martin, 2002; Green & Bohanann, this volume). Because of differences in rate of molec-ular evolution among species in most groups of organisms, dissimilarity in molecular sequences is not a consistent measure of evolutionary divergence across taxa. Divergence time estimation from DNA sequence data often requires calibration points using independent evidence, such as dates of unambiguous evolutionary splits inferred from fossil or biogeographic data (Arbogast *et al.*, 2002). Thus, in the present work, we shall interpret the term $t_{ij}$ as a divergence time, assuming problems associated with variance in rate of nucleotide substitutions have been dealt with using appropriate approaches (e.g. Sanderson, 1997, 2002). The term $t_{ij}$ hence denotes the age of the most recent common ancestor of taxon $i$ and $j$, divided by the age of the most recent common ancestor of the whole set of taxa; that is, $t_{ij}$ is between 0 and 1. A simple measure of among-locality differentiation (Nei, 1973; Slatkin, 1991) can be defined as

$$F_{ST} = \frac{D_\beta}{D_T}, \tag{8.7}$$

where this variable is defined by analogy to the $F_{ST}$ in population genetics and measures community turnover instead of population genetic differentiation.

### Data and statistical analyses

To test the relevance of our approach we used a large tropical plant census collected by A. H. Gentry (Gentry, 1982, 1988; Phillips & Miller, 2002), which was made available online by the Missouri Botanical Garden (www.mobot.org/MOBOT/Research/gentry/transect.shtml). For the purpose of this work, we used only the data for the 124 South American sites. Each site consists of 10 randomly placed transects $2 \times 50$ m in size, totaling roughly 0.1 ha in area. In these plots all woody ferns, trees, lianas and monocots above 2.5 cm in diameter (originally, 1 inch) were identified to the species, or sorted into morphospecies (*c.* 95% of the individuals were identified at least at the species or genus level, see below). For

each site, an elevation was recorded (range 10–3050 m a.s.l.) as well as annual rainfall for 123 of the 124 sites (range 400–9000 mm/yr).

The data set contained a total of 44 114 stems. To minimize the number of spelling errors in the data set we matched the genus names to a list of all plant genera maintained by the Royal Botanic Gardens at Kew (www.rbgkew.org.uk/web.dbs/genlist.html), and corrected typos. We then verified that the family names matched the most recent molecular-based plant phylogeny, based on the work of the Angiosperm Phylogeny Group (APG, 2003), and on Soltis *et al.* (2002) for more ancestral nodes (tracheophytes). In total, the data set we used contained 4471 morphospecies. Of these, 1094 taxa were described only to the genus level (38.7% of the individuals) and 56 to the family level (6.2% of the individuals). For instance, 732 individuals were grouped into 16 morphospecies of the difficult genus *Miconia* (Melastomataceae). Only 115 individuals could not be identified even to the family level (0.2%). It is important to emphasize that although many papers have made use of the Gentry data set, not all have achieved the same effort of data standardization. Also, many voucher specimens of the Gentry collection are continuously being identified by the specialists of the Missouri Botanical Garden. Thanks to their dedication, the data set used here is considerably improved over that used a decade ago (for an overview, see, e.g. Phillips & Miller, 2002).

Next, we obtained a phylogeny for the species included in this data set by using a method inspired from that of Webb (2000). A super-tree containing the full APG II phylogeny (APG, 2003), plus other literature sources, were assembled by C. O. Webb (Webb & Donoghue, 2004; www.phylodiversity.net), updated for the present work, and pruned accordingly. In the tree, terminal nodes always correspond to species, branching off from genus-level nodes. The most ancestral node corresponds to the origin of vascular plants (euphyllophytes). We emphasize that our phylogenetic hypothesis has many issues, being unevenly resolved across clades, being the result on a "grafting" of the phylogeny of selected families onto the APG phylogeny. Thus our quantitative results are likely to be altered when better phylogenetic hypotheses are made available. However, we are confident that the present hypothesis faithfully reflects the current stage of knowledge in plant phylogenetics.

Finally, we "dated" the phylogeny using a calibration produced by Wikström *et al.* (2001), who combined fossil data and existing molecular-based data on angiosperms to date divergence times for over 400 interior nodes, mostly families and orders. Nodes above the family level used all angiosperm families, so we could define the age of the nodes as the maximal age between two taxa within the clade. The family level nodes were more difficult to resolve, as they were based on a small subsample of the total number of genera. For instance, if only two genera of a genus-rich family were included in the phylogeny, then the divergence time between these two genera would likely underestimate the true

age of the family. We further discuss this issue in the Appendix. We postulated an age of 450 Myr for the origin of euphyllophytes (the most ancestral node in our phylogeny), and 166 Myr for the origin of Cyatheaceae (tree ferns). Soltis *et al*. (2002) showed that divergence times of tree ferns tend to be underestimated using molecular-based techniques, so we privileged the estimate based on fossil information. This tree enabled us to construct a matrix $t_{ij}$, representing divergence times between species $i$ and species $j$. For instance, the cucumber family and the beech family diverged 84 Myr ago, so the relative divergence time between *Cucurbita maxima* and *Fagus sylvatica* would be $84/450 = 0.187$. The divergence time of any two distinct genera within the same family was estimated by the divergence time of the family, as inferred in the Appendix. For instance, the divergence between *Iryanthera sagotiana* and *Virola calophylla*, both in the Myristicaceae, was taken to be 46 Myr, so $t_{ij} = 46/450 = 0.1$. Because the species level was not resolved in our phylogenetic hypothesis, we assumed that the divergence between two congeneric species was equal to the estimated genus age.

We contrasted results obtained by using this distance matrix, to a star-shaped phylogeny, such that $t_{ij} = 1$ for all $i \neq j$, and $t_{ij} = 0$, from which we computed the nonphylogenetic diversity indices $\overline{D}_\alpha, \overline{D}_\beta$ (we remind the reader that nonphylogenetic measures of diversity are distinguished from phylogenetic measures by the bar over the symbol).

First, we tested whether the addition of a phylogenetic structure significantly modified measures of biodiversity. A simple approach to this question is to ask whether the alpha diversity, as measured by $D_\alpha$, and the among-site differentiation, as measured by $F_{ST}$, significantly differ from a null expectation. A direct comparison between the phylogenetic measure $D$ and the nonphylogenetic measure $\overline{D}$ is not straightforward given the different scaling of times (however, the dimensionless $F_{ST}$ may be compared). Instead, we constructed a randomization test by which we kept the species lists unchanged in the sites, but reshuffled the species location across the phylogeny (Slatkin, 1991). Hence, the tree structure remains unchanged: only the labels of the tip nodes are permuted. For each randomized phylogeny, we computed $D_\alpha$ and $F_{ST}$, and we compared the observed value with the null expectation by means of a simple Student's *t*-test.

To test for spatial variation in diversity, we compared the diversity matrices with environmental or geographical distance matrices, using the Mantel test. The Mantel test compares two similarity or distance matrices computed for the same sites (Legendre & Legendre, 1998). The Mantel statistic $r_M$ is a measure of the correlation between the two matrices, and behaves like a correlation coefficient. It is usually tested by a nonparametric permutational test as the assumption of independence between values of the variable is not fulfilled in similarity or distance matrices.
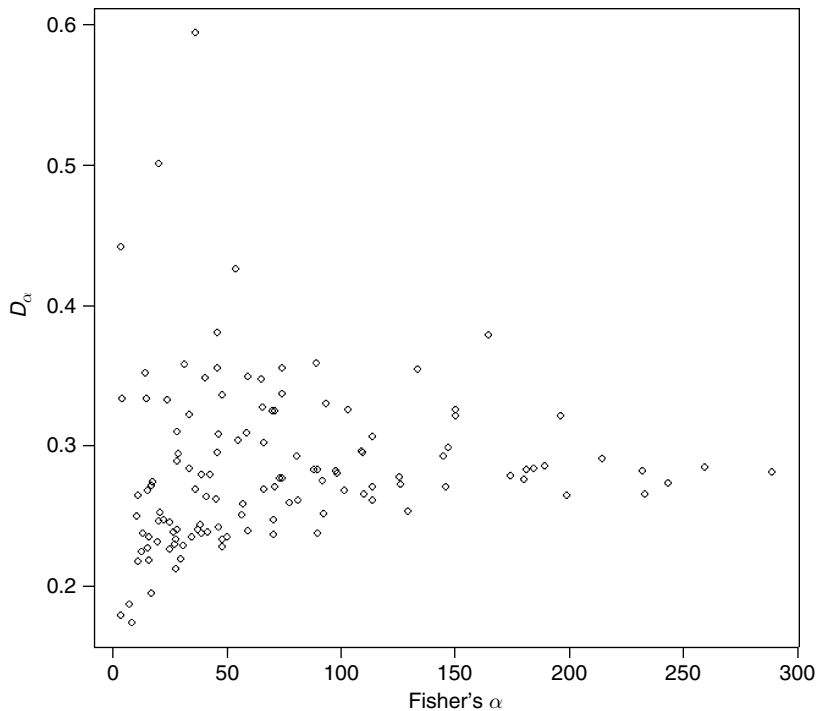
**Figure 8.1**  Scatter plot of $D_{\alpha}$ against Fisher's $\alpha$, a measure of local species diversity controlling for sample size.

## Results

First, we asked if nonphylogenetic diversity is a suitable approximation of phylogenetic diversity. Figure 8.1 illustrates a comparison between the measure of phylogenetic alpha diversity $D_{\alpha}$ and a commonly used measure of alpha diversity, Fisher's alpha (see e.g. ter Steege *et al.*, 2003). Except at low diversity values, there was no significant relationship between these two measures: the plots with highest species diversity were not necessarily those with largest phylogenetic diversity. We also compared the phylogenetic and nonphylogenetic measures of beta diversity. Specifically, we performed a Mantel test to compare the phylogenetic diversity matrix $H$ with the nonphylogenetic diversity matrix $\overline{H}$. Mantel's correlation coefficient was found to be $r_M = 0.66$, and the significance level, obtained by permutations was $P = 0.001$. Thus, nonphylogenetic beta diversity is significantly correlated with phylogenetic beta diversity.

Second, we compared the mean phylogenetic diversity to that measured with randomized phylogenetic trees. In so doing, we also tested for the presence of phylogenetic structure in the data. Mean local phylogenetic diversity and phylogenetic beta diversity were equal to
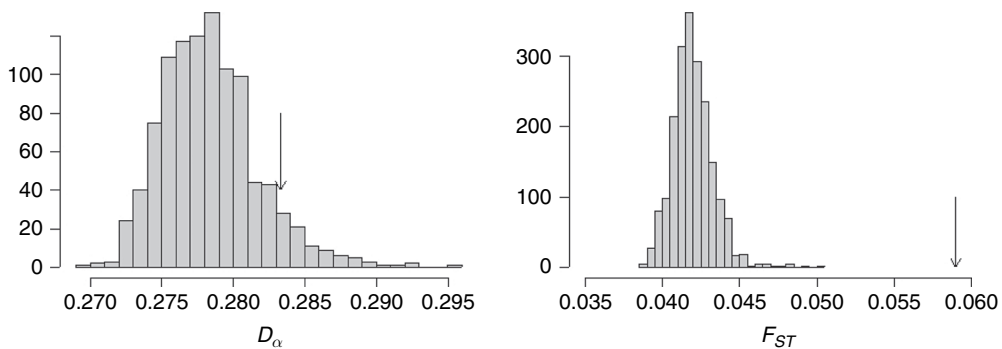
$$D_{\alpha} = 0.283, F_{ST} = 0.059.$$

**Figure 8.2** Values of local (alpha) and intersite (beta) phylogenetic diversity as measured by $D_\alpha$ (left) and $F_{ST}$ (right) for the Gentry data set (arrows). Histograms are the randomized measures of phylogenetic diversity (based on 1000 replicates). The intersite diversity as measured by the $F_{ST}$ is significantly higher in the Gentry data set than in values obtained from the randomization test.

These average figures were compared with the randomized values:

$$D_\alpha|_{rand} = 0.278, F_{ST}|_{rand} = 0.042.$$

Thus, average local phylogenetic diversity is not significantly modified by a randomization in the phylogenetic structure ($t = 0.62$, $P > 0.5$). However, phylogenetic beta diversity as measured by the $F_{ST}$ does differ from the null expectation ($t = 12.6$, $P < 0.001$). This test is further illustrated in Fig. 8.2. Cross-plot local phylogenetic diversity $D_\alpha$ varied from 0.17 for Puyeyhue (Ecuador) to 0.59 for Alto de Mirador (Colombia) while the randomized measure of local diversity varied from 0.19 to 0.37. The randomization test summarized in Fig. 8.3 also shows that 26 plots are significantly less phylogenetically diverse locally than the null expectation, and 24 plots are more phylogenetically diverse (*t*-test, $P < 0.05$). Measures of phylogenetic beta diversity were also tested individually: 24% of the values were significantly greater than the randomized value, while 14% were significantly smaller (*t*-test, $P < 0.05$). Two sites were outliers in this analysis, having very high local phylogenetic diversities: Alto de Mirador ($D_\alpha = 0.59$) and Alto de Sapa ($D_\alpha = 0.50$), both from the highlands of Colombia. This is explained by the abundance of ancient plant groups that were rare in the other plots (e.g. tree ferns, *Cyathea* spp.). Excluding these two sites, however, we found $D_\alpha = 0.279$, $F_{ST} = 0.055$, and the significance tests for $D_\alpha$ and $F_{ST}$ were not modified. Hence, our results are not driven by outliers.

Third, we used a correlative approach to relate observed patterns of phylogenetic alpha and beta diversity. For alpha diversity, we used a linear model to correlate $D_\alpha$ with elevation and rainfall. Both rainfall and (rainfall)^2 were significant predictors of local diversity, and they explained 22% of the total variance. This contrasts with the nonphylogenetic measure of local diversity $\overline{D}_\alpha$ in which we found only a weak signal: only rainfall was significant

**Table 8.1** *Linear regression of phylogenetic and nonphylogenetic measures of local (alpha) diversity against elevation, and annual rainfall*
Numbers represent the partial $R^2$ for the variables. We also report the best regression model.

|  | Phylogenetic local diversity $D(A, A)$ | Nonphylogenetic local diversity $\bar{D}(A, A)$ |
|---|---|---|
| Elevation | 0.127, $P < 10^{-4}$ | 0.0282, $P = 0.062$ |
| Rainfall | 0.151, $p < 10^{-4}$ | 0.0708, $P = 0.003$ |
| (Rainfall)^2 | 0.075, $P = 0.001$ | 0.0178, $P = 0.129$ |
| Variables of regression model | Rainfall, (Rainfall)^2 | Rainfall |
| Predicted $R^2$ for regression model | $R^2 = 0.226$ | $R^2 = 0.0708$ |

**Table 8.2** *Mantel correlations ($r_M$) and significance levels (based on 999 permutations) for both phylogenetic and nonphylogenetic diversity matrices H (a measure of beta diversity), compared with geographical distance (GD), logarithmically transformed geographical distance, elevation, and annual rainfall*

|  | Phylogenetic diversity $H(A, B)$ | Nonphylogenetic diversity $\bar{H}(A, B)$ |
|---|---|---|
| $\bar{H}(A, B)$ | 0.6604, $P = 0.001$ |  |
| Geographical distance (*GD*) | 0.2606, $P = 0.001$ | 0.3015, $P = 0.001$ |
| ln(*GD*) | 0.1908, $P = 0.001$ | 0.2144, $P = 0.001$ |
| Elevation | 0.2264, $P = 0.001$ | 0.1273, $P = 0.037$ |
| Annual rainfall | 0.0026, $P = 0.371$ | 0.0458, $P = 0.491$ |
| Variables of regression model | Elevation, *GD* | Elevation, *GD* |
| Predicted $R^2$ for regression model | $R^2 = 0.133$ | $R^2 = 0.131$ |

(Table 8.1). For the phylogenetic and nonphylogenetics beta-diversity matrices $H$ and $\overline{H}$, we used Mantel tests. We correlated them with environmental and geographical distance matrices. The results are displayed in Table 8.2. Geographical distance was the best descriptor of both phylogenetic and non-phylogenetic diversity ($r_M = 0.261$, and $r_M = 0.301$, respectively). Elevation was also a significant predictor in both cases, but phylogenetic diversity was better predicted by this variable than nonphylogenetic diversity ($r_M = 0.226$, and $r_M = 0.127$, respectively).

## Discussion

We showed that incorporating phylogenetic information affects both the local and regional biodiversity patterns observed in neotropical plant communities. We detected a strong phylogenetic signal on estimation of mean beta-diversity
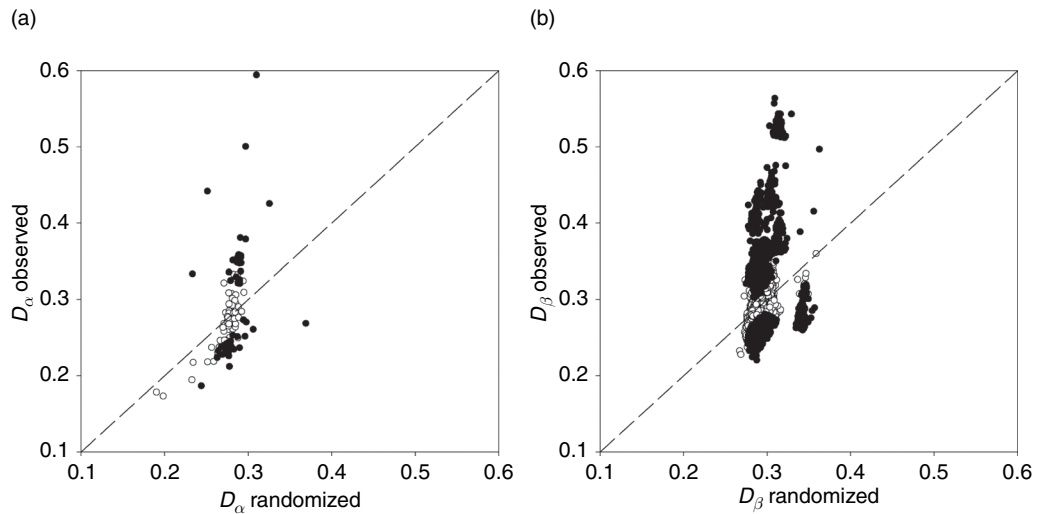
**Figure 8.3** (a) Observed values of $D_\alpha$ for the 124 Gentry plots, against the null expectation (mean $D_\alpha$ obtained with the randomization test). (b) Observed values of $D_\beta$ for pairs of Gentry plots plotted against the randomized value. Both panels evidence the large range in diversity within and across plots. Black dots refer to values significantly lower or higher than the null expectation.

$F_{ST}$; a site-by-site analysis, as reported in Fig. 8.3, also revealed a strong local phylogenetic signal in 50 sites out of 124. This shows that phylogenies encapsulate an information that is not summarized in species-based measures of diversity, both for local diversity (at least in a large fraction of the sites) and for beta diversity. In other words, species diversity does not accurately reflect overall evolutionary diversity, a central message of the present work.

Why are some sites significantly more phylogenetically diverse than expected by chance? A closer inspection at the species assemblages in these sites reveals that they had a greal deal of species belonging to ancient clades, like tree ferns or basal dicot species, in the Magnoliales and in the Laurales. They were predominantly montane species, which supports the often discussed biogeographic pattern of an origin of the angiosperms in the montane tropics (Axelrod, 1952; Takhtajan, 1969). Other sites were less phylogenetically diverse than expected by chance, due to the overrepresentation of single families or orders. This high degree of phylogenetic clustering is compatible with two alternative ecological mechanisms. It could either be that sites are overrepresented by a single species, which might be a poor disperser. This clumping effect would then be an indirect evidence for dispersal limitation in tropical floras. Alternatively, species of most clades but one might have been filtered out by the environment, suggesting that the species co-occurring in the site would have been selected for a biological feature not shared by the species of other clades.

For example, the hyperabundance of palm species in neotropical swamps prob-
ably reflects a special adaptation of this clade to the anaerobic growth of the root
system. It would be interesting to explore further such patterns of phylogenetic
clustering or overdispersion over large spatial scales. Notably, besides elevation,
only geographical distance stood out as a significant predictive factor of the
phylogenetic diversity matrix $H$.

So far, conservation planning at the regional scale has almost exclusively
focused on species count data. We have shown here that for neotropical plants,
species richness does not predict phylogenetic diversity. While Prance (1994)
found that diversity above the species level exhibited considerably smoother
variation than species level patterns in neotropical plant communities, our
results suggest that reservoirs for phylogenetic history cannot always be
deduced from patterns of local species richness. Thus, it is clearly important
to incorporate phylogenetic information, when available, in planning conser-
vation strategies or in identifying areas of high biodiversity. Also, it should be
pointed out here that consideration of phylogenetic structure does not circum-
vent the problem of first documenting species diversity. In view of the limi-
tations of such protocols, discussed in the Introduction, one might wonder
why this refinement is at all a useful step forward. A first practical answer to
this question is that, within the formalism of phylogenetic diversity developed
above, one can easily accommodate heterogeneous taxonomic censuses. For
instance, if a fraction of the taxa are known only to the family level or to the
genus level, it is still possible to explore their relatedness to other taxa. Thus
measures of phylogenetic diversity are less sensitive to the issue of sorting taxa
into morphospecies. Further, phylogenetic diversity has a natural connection
to the rapidly expanding field of DNA bar coding. This technique largely
bypasses classical taxonomic identification by sequencing universal genes,
which are sufficiently variable to show a significant level of interspecific
signal (Hebert *et al.*, 2004; see however Moritz & Cicero, 2004). While unable
to resolve deep nodes of the phylogeny, such markers could be used to develop
an individual-based phylogeny of an ecological community, within which
species would be considered as monophyletic clades. This would partially
circumvent the problem of taxonomic identification and at the same time
would provide a method for assessing the total amount of divergence within
an ecological community.

Our analysis is a first attempt to relate concepts of evolutionary diversity and
empirical data both for alpha diversity and beta diversity. While quantifying
biodiversity using phylogenetic information is, at least conceptually, an obvious
improvement over traditional approaches, the present attempt still suffers from
a number of limitations. Despite our efforts to use well-resolved phylogenies
and to calibrate the resulting super-tree, our estimates of evolutionary diversity
are not devoid of possible bias. In the case of flowering plants, a number of

important nodes remain poorly supported, and the branch lengths as well as the tree topology remain indicative, particularly below the family level (see Saladin *et al*., 2005, for recent developments). For instance, in one important order for neotropical woody plant species, the Malpighiales, it has been shown that Wikström *et al*. (2001)'s estimate of the ages in this clade was considerably underestimated (Davis *et al*., 2005), a common situation when fossil evidence is limited or controversial. It may be argued that diversity measures based on nucleotide differences (the basis for the construction of phylogenetic trees), rather than on divergence times would produce estimates that would be less dependent on assumptions on diversification rates. However, we contend that our measure is more biologically relevant, as it directly relates to the evolutionary history of the species assemblages under study. For instance, our result that the minimum local diversity is $D_\alpha = 0.17$ may be reinterpreted as follows: on average two randomly chosen individuals in the community come from lineages that diverged 76.5 Myr ago. Clarke and Warwick, in a series of recent reports (Clarke & Warwick, 2001; Warwick & Clarke, 2001) defined very similar measures of phylogenetic diversity, in which branch lengths measure taxonomic levels (species, genus, family, suborder, order, and so forth). This measure of "taxonomic diversity" gives an equal weight to recent and to old lineages (e.g. the fairly recent Chrysobalanaceae and the very ancient Piperaceae), but should provide a good approximation to phylogenetic diversity in cases where no reliable phylogeny is available, or for historical species lists, as pointed out by Clarke and Warwick (2001).

## Acknowledgments

## Appendix 8.I   Maximum likelihood estimation of the age of a family from an incomplete sampling

For each family, Wikström *et al*. (2001) listed the maximal possible age, $T_{max}$ (divergence time to the sister family), the minimal age $T_{min}$ (age of the most recent ancestral node if more than one genus is present; zero otherwise), $k$ the number of sampled genera. Moreover, $K$, the total number of genera in each family, is available. An estimate of $T$, the true age of the family, is a function of these four parameters:

$$T = f(T_{min}, T_{max}, k, K).$$

**Table 8.3** *Estimation of the family age for the most represented families in the Gentry data set*
$T_{min}$ is the age of the most ancestral genus in the family, $T_{max}$ the divergence time to the sister family, $k$ the number of sampled genera within the family. All three parameters are reported in Wikström *et al*. (2001). $K$ is the total number of extant genera in this family after the Kew checklist of plant genera (in parentheses, after Smith *et al*. 2004), and $N$, the number of sampled individuals in the Gentry data set. The sampling parameter $\beta$ is estimated from Eq. (8.1) if $k > 1$ or set to 1 if $k = 1$ (uniform prior sampling distribution). $T$ is the estimated family age.

| Family name | $N$ | $T_{min}$ | $T_{max}$ | $k$ | $K$ | $\beta$ | $T$ |
|---|---|---|---|---|---|---|---|
| Annonaceae | 954 | 63 | 82 | 2 | 125(135) | 0.98 | 81 |
| Apocynaceae | 732 | 18 | 45 | 2 | 496(250–550) | 2.82 | 41 |
| Araceae | 547 | 98 | 124 | 2 | 109(105) | 0.92 | 123 |
| Arecaceae | 2387 | 73 | 99 | 7 | 205(189) | 1.81 | 94 |
| Asteraceae | 588 | 44 | 50 | 5 | 1511(1535) | 1.009 | 50 |
| Bignoniaceae | 2024 | 38 | 47 | 3 | 111(120) | 1.05 | 46 |
| Burseraceae | 571 | 0 | 51 | 1 | 18(18) | 1 | 48 |
| Celastraceae | 522 | 42 | 58 | 5 | 92(76) | 1.66 | 55 |
| Clusiaceae | 917 | 0 | 45 | 1 | 47(36) | 1 | 44 |
| Euphorbiaceae | 1670 | 0 | 69 | 1 | 333(300) | 1 | 69 |
| Fabaceae | 3762 | 56 | 79 | 3 | 677(650–700) | 1.41 | 78 |
| Lauraceae | 1397 | 34 | 80 | 2 | 49(52) | 2.61 | 64 |
| Lecythidaceae | 530 | 65 | 88 | 3 | 25(20) | 1.3 | 81 |
| Malvaceae | 876 | 34 | 54 | 8 | 273(245)[a] | 2.54 | 49 |
| Melastomataceae | 1691 | 0 | 41 | 1 | 194(155) | 1 | 41 |
| Meliaceae | 1088 | 30 | 40 | 2 | 51(50) | 1.04 | 39 |
| Moraceae | 1474 | 23 | 36 | 2 | 37(37) | 1.42 | 33 |
| Myristicaceae | 840 | 23 | 113 | 2 | 18(19) | 6.35 | 46 |
| Polygonaceae | 536 | 26 | 37 | 2 | 49(45) | 1.19 | 36 |
| Rubiaceae | 2844 | 56 | 64 | 4 | 623(650) | 0.83 | 64 |
| Salicaceae | 646 | 40 | 53 | 5 | 82(–) | 1.52 | 50 |
| Sapindaceae | 1020 | 36 | 56 | 5 | 138(147) | 2.71 | 48 |
| Sapotaceae | 860 | 0 | 54 | 1 | 59(53) | 1 | 53 |
| Urticaceae | 503 | 22 | 42 | 2 | 55(46–58)[a] | 1.95 | 37 |
| Violaceae | 642 | 0 | 51 | 1 | 20(25) | 1 | 48 |

[a] Sum of number of genera considered as different families in Smith *et al*. (2004)'s treatment.

We constructed the following probabilistic model. By definition, the age $T$ of a family is the maximal age of all genera within the family. Let $\{x_1, x_2, \Lambda, x_K\}$ be the sequence of ages of the $K$ genera within a family. Then:

$$T = \max_{i \leq K}(x_i),$$

where, for all $i$, $x_i < T_{max}$.

The simplest model consists in assuming that the genus ages are Poisson-distributed between 0 and $T_{\max}$, and to estimate $T$ for $K$ draws using extreme value theory. However, we already know that $k$ genera were sampled, out of the total of $N$ genera. Let us assume that the genus ages $x_i$ are distributed according to a beta probability density function $P(x) = A(T_{\max} - x)^{\beta - 1}$, with $0 < x < T_{\max}$, and $A = \beta/T_{\max}^{\beta}$. If $\beta > 1$, values of $x$ close to $T_{\max}$ are less likely, hence the situation corresponds to a family with genera that all tend to be younger than $T_{\max}$. The extreme value probability density function of this process is given by

$$F_k(x) = kP(x)\left(\int_0^x P(y)dy\right)^{k-1}.$$

This is the probability that in $k$ trials, all $k$ but one are below the value $x$. In our case, we define a likelihood function

$$L(\beta|k, T_{\min}, T_{\max}) = F_k(T_{\min}).$$

The maximal likelihood estimator of $\beta$ is defined implicitly by

$$\beta\left(1 - \frac{k-1}{(1-\tau)^{-\beta}-1}\right) + \frac{1}{\ln(1-\tau)} = 0, \tag{8.8}$$

with $\tau = \frac{T_{\min}}{T_{\max}}$.

The expected age of the family can be estimated as the first moment of $F_K(x)$:

$$T = T_{\max}\left[1 - \int_0^1 \left(1 - x^\beta\right)^K dx\right] = T_{\max}\left[1 - \frac{B(K+1, 1/\beta)}{\beta}\right]. \tag{8.9}$$

For each family, we first computed parameter $\beta$ as a function of $k$ and $\tau$ by solving iteratively Eq. (8.8). Then we used this parameter together with $K$, the total number of genera in the family, to provide an estimate of the family age. This procedure is illustrated in Table 8.3 for the most abundant families in the Gentry data set.

## References

Alroy, J. (2002). How many named species are valid? *Proceedings of the National Academy of Sciences of the USA*, **99**, 3706–3711.

Angiosperm Phylogeny Group [APG]. (2003). An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG II. *Botanical Journal of the Linnean Society*, **141**, 399–436.

Arbogast, B. S., Edwards, S. V., Wakeley, J., Beerli, P. & Slowinksi, J. J. (2002). Estimating divergence times from molecular data on phylogenetic and population genetic time-scales. *Annual Review of Ecology and Systematics*, **33**, 707–740.

Axelrod, D. I. (1952). A theory of angiosperm evolution. *Evolution*, **4**, 29–60.

Brown, J. H. (1995). *Macroecology*. Chicago: University of Chicago Press.

Clark, D. A., Clark, D. B., Sandoval, R. M. & Castro Vinicio, M. C. (1995). Edaphic and human effects on landscape-scale distributions of

tropical rain forest palms. *Ecology*, **76**, 2581–2594.

Clarke, K. R. & Warwick, R. M. (2001). A further biodiversity index applicable to species lists: variation in taxonomic distinctness. *Marine Ecology Progress Series*, **216**, 265–278.

Davis, C. C., Webb, C. O., Wurdack, K. J., Jaramillo, C. A. & Donoghue, M. J. (2005). Explosive radiation of Malpighiales supports a mid-Cretaceous origin of modern tropical rain forests. *American Naturalist*, **165**, E36–E65.

Erwin, T. L. (1982). Tropical forests: their richness in coleopteran and other arthropod species. *Coleopterists Bulletin*, **36**, 74–75.

Ewens, W. J. (1972). The sampling theory of selectively neutral alleles. *Theoretical Population Biology*, **3**, 87–112.

Faith, D. P. (1994). Phylogenetic pattern and the quantification of organismal diversity. *Philosophical Transactions of the Royal Society of London*, *Series B*, **345**, 45–58.

Gentry, A. H. (1982). Patterns of Neotropical plant species diversity. *Evolutionary Biology*, **15**, 1–84.

Gentry, A. H. (1988). Changes in plant community diversity and floristic composition on environmental and geographical gradients. *Annals of the Missouri Botanical Garden*, **75**, 1–34.

Gotelli, N. J. & Colwell, R. K. (2001). Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. *Ecology Letters*, **4**, 379–391.

Hebert, P. D., Stoeckle, M. Y., Zemlak, T. S. & Francis, C. M. (2004). Identification of birds through DNA barcodes. *PLoS Biology*, **2-e**, 312.

Hubbell, S. P. (2001). *A Unified Neutral Theory of Biodiversity and Biogeography*. Princeton, NJ: Princeton University Press.

Kahn, F. & Mejia, K. (1991). The palm communities of two "terra firme" forests in Peruvian Amazonia. *Principes*, **35**, 22–26.

Krebs, C. J. (1998). *Ecological Methodology*, 2nd Edn. Reading, MA: Addison-Welsey.

Lande, R. (1996). Statistics and partitioning of species diversity, and similarity among multiple communities. *Oikos*, **76**, 5–13.

Legendre, P. & Legendre, L. (1998). *Numerical Ecology*, 2nd English edition. Amsterdam: Elsevier Science.

Magurran, A. E. (1988). *Ecological Diversity and its Measurement*. London: Chapman and Hall.

Martin, A. P. (2002). Phylogenetic approaches for describing and comparing microbial communities. *Applied and Environmental Microbiology*, **68**, 3673–3682.

May, R. M. (1975). Patterns of species abundance and diversity. In *Ecology and Evolution of Communities*, ed. M. L. Cody & J. M. Diamond, pp. 81–120. Cambridge, MA: Belknap Press of Harvard University Press.

May, R. M. (1994). Conceptual aspects of the quantification of the extent of biological diversity. *Philosophical Transactions of the Royal Society of London*, *Series B*, **345**, 13–20.

Mooers, A. O. & Heard, S. B. (1997). Inferring evolutionary process from phylogenetic tree shape. *Quarterly Review of Biology*, **72**, 31–54.

Moritz, C. & Cicero, C. (2004). DNA barcoding: promise and pitfalls. *PLoS Biology*, **2-e**, 279.

Nee, S. & May, R. M. (1997). Extinction and the loss of evolutionary history. *Science*, **278**, 692–694.

Nei, M. (1973). Analysis of gene diversity in subdivided populations. *Proceedings of the National Academy of Sciences of the USA*, **70**, 3321–3323.

Nei, M. (1987). *Molecular Evolutionary Genetics*. New York: Columbia University Press.

Novotný, V., Basset, Y., Miller, S. E., *et al.* (2002). Low host specificity of herbivorous insects in a tropical forest. *Nature*, **416**, 841–844.

Pavoine, S., Ollier, S. & Dufour, A. B. (2005). Is the originality of a species measurable? *Ecology Letters*, **8**, 579–586.

Phillips, O. L. & Miller, J. (2002). *Global Patterns of Plant Diversity: Alwyn H. Gentry's Forest Transect Data Set*. St. Louis, MO: Missouri Botanical Garden.

Prance, G. T. (1994). A comparison of the efficacy of higher taxa and species numbers in the assessment of biodiversity in the Neotropics. *Philosophical Transactions of the Royal Society of London*, *Series B*, **345**, 89–99.

Ricklefs, R. E. (2004). A comprehensive framework for global patterns in biodiversity. *Ecology Letters*, **7**, 1–15.

Ricklefs, R. E. & Schluter, D. (eds) (1993). *Species Diversity in Ecological Communities: Historical and Geographical Perspectives*. Chicago: University of Chicago Press.

Saladin, N., Hodkinson, T. R. & Savolainen, V. (2005). Towards building the tree of life: a simulation study for all angiosperm genera. *Systematic Biology*, **54**, 183–196.

Sanderson, M. J. (1997). A nonparametric approach to estimating divergence times in the absence of rate constancy. *Molecular Biology and Evolution*, **14**, 1218–1231.

Sanderson, M. J. (2002). Estimating absolute rates of molecular evolution and divergence times: a penalized likelihood approach. *Molecular Biology and Evolution*, **19**, 101.

Smith, N., Mori, S. A., Henderson, A., Stevenson, D. W. & Heald, S. V. (2004). *Flowering plants of the Neotropics*. Princeton University Press, pp. 594.

Soltis, P. S., Soltis, D. E., Savolainen, V., Crane, P. E. & Barraclough, T. G. (2002). Rate heterogeneity among lineages of tracheophytes: integration of molecular and fossil data and evidence for molecular living fossils. *Proceedings of the National Academy of Sciences of the USA*, **99**, 4430–4435.

Slatkin, M. (1991). Inbreeding coefficients and coalescent times. *Genetic Research*, **58**, 167–175.

Suau, A., Bonnet, R., Sutren, M., *et al.* (1999). Direct analysis of genes encoding 16S rRNA from complex communities reveals many novel molecular species within the human gut. *Applied and Environmental Microbiology*, **65**, 4799–4807.

Takhtajan, A. (1969). *Flowering Plants: Origin and Dispersal*. Edinburgh: Oliver & Boyd.

ter Steege, H., Pitman, N., Sabatier, D., *et al.* (2003). A spatial model of tree alpha-diversity and tree density for the Amazon. *Biodiversity and Conservation*, **12**, 2255–2277.

Tuomisto, H. & Poulsen, A. D. (2000). Pteridophyte diversity and species composition in four Amazonian rain forests. *Journal of Vegetation Science*, **11**, 383–396.

Tuomisto, H., Ruokolainen, K. & Yli-Halla, M. (2003). Dispersal, environment, and floristic variation of Western Amazonian forests. *Science*, **299**, 241–244.

Vane-Wright, R. I., Humphries, C. J. & Williams, P. H. (1991). What to protect? Systematics and the agony of choice. *Biological Conservation*, **55**, 235–254.

Vormisto, J., Svenning, J. C., Hall, P. & Balslev, H. (2004). Diversity and dominance in palm (Arecaceae) communities in terra firme forests in the western Amazon basin. *Journal of Ecology*, **92**, 577–588.

Warwick, R. M. & Clarke, K. R. (2001). Practical measures of marine biodiversity based on relatedness of species. *Oceanography and Marine Biology*, **39**, 207–231.

Watterson, G. A. (1974). Models for the logarithmic species abundance distributions. *Theoretical Population Biology*, **6**, 217–250.

Webb, C. O. (2000). Exploring the phylogenetic structure of ecological communities: an example for rain forest trees. *American Naturalist*, **156**, 145–155.

Webb, C. O. & Donoghue, M. J. (2004). Phylomatic: tree retrieval for applied phylogenetics. *Molecular Ecology Notes*, **5**, 181–183.

Webb, C. O. & Pitman N. C. A. (2002). Phylogenetic balance and ecological evenness. *Systematic Biology*, **51**, 898–907.

Webb, C. O., Ackerly, D. D., McPeek, M. A. & Donoghue, M. J. (2002). Phylogenies and community ecology. *Annual Review of Ecology and Systematics*, **33**, 475–505.

Wikström, N., Savolainen, V. & Chase, M. W. (2001). Evolution of the angiosperms: calibrating the family tree. *Proceedings of the Royal Society of London, Series B*, **268**, 2211–2220.

Whittaker, R. H. (1972). Evolution and measurement of species diversity. *Taxon*, **21**, 213–251.

Wright, S. (1931). Evolution in Mendelian populations. *Genetics*, **16**, 97–159.

Yule, G. U. (1925). A mathematical theory of evolution based on the conclusions of Dr J. C. Willis. *Philosophical Transactions of the Royal Society of London, Series B*, **213**, 21–87.