Beta diversity as the variance of community data: dissimilarity coefficients and partitioning

Pierre Legendre[1*], Miquel De Cáceres[2,3]

[1] *Département de sciences biologiques, Université de Montréal, C.P. 6128, succursale Centre-ville, Montréal, Québec, Canada H3C 3J7*

[2] *Centre Tecnològic Forestal de Catalunya. Ctra. St. Llorenç de Morunys km 2, 25280 Solsona, Catalonia, Spain*

[3] *CREAF (Centre de Recerca Ecològica i Aplicacions Forestals), Bellaterra, Catalonia, Spain*

E-mail addresses: pierre.legendre@umontreal.ca, miquelcaceres@gmail.com

Short running title: Beta diversity partitioning

Keywords: beta diversity, community ecology, community composition data, dissimilarity coefficients, local contributions to beta diversity, properties of dissimilarity coefficients, species contributions to beta diversity, variance partitioning

Article type: Ideas and Perspectives

# Words in abstract: 200	# Words in main text: 7500	# Text boxes: 0
# References: 61	# Figures: 4	# Tables: 2

*Correspondence
Pierre Legendre
Département de sciences biologiques
Université de Montréal
C.P. 6128, succursale Centre-ville
Montréal, Québec, Canada H3C 3J7

Phone: 514-343-7591
Fax: 514-343-2293
E-mail: Pierre.Legendre@umontreal.ca

Authorship statement

The two authors contributed equally to the paper and took the lead at different times. PL coordinated the writing and editing of the final version of the manuscript.

Abstract

Beta diversity can be measured in different ways. Among these, the total variance of the community data table **Y** can be used as an estimate of beta diversity. We show how the total variance of **Y** can be calculated either directly or through a dissimilarity matrix obtained using any dissimilarity index deemed appropriate for pairwise comparisons of community composition data. We addressed the question of which index to use by coding 16 indices using 14 properties that are necessary for beta assessment, comparability among data sets, sampling issues, and ordination. Our comparison analysis classified the coefficients under study into five types, three of which are appropriate for beta diversity assessment. Our approach links the concept of beta diversity with the analysis of community data by commonly used methods like ordination and ANOVA. Total beta can be partitioned into Species Contributions (SCBD: degree of variation of individual species across the study area) and Local Contributions (LCBD: comparative indicators of the ecological uniqueness of the sites) to Beta Diversity. Moreover, total beta can be broken up into within- and among-group components by MANOVA, into orthogonal axes by ordination, into spatial scales by eigenfunction analysis, or among explanatory data sets by variation partitioning.

INTRODUCTION

A most interesting property of species diversity is its organization through space. This phenomenon, now well known to community ecologists, was first discussed by Whittaker in two seminal papers (1960, 1972) where he described the alpha, beta and gamma diversity levels of natural communities. Alpha is local diversity, beta is spatial differentiation, and gamma is regional diversity. The interest of community ecologists for beta diversity stems from the fact that spatial variation in species composition allows them to test hypotheses about the processes that generate and maintain biodiversity in ecosystems. Sampling through space, time, or along gradients representing processes of interest is a way of carrying out *mensurative experiments* (Hurlbert 1984) involving natural processes without the constraints (e.g. small sample size) of controlled experiments.

Beta diversity is conceptually the variation in species composition among sites within a geographic area of interest (Whittaker 1960). Several authors have used that description of the concept, including Legendre *et al.* (2005), Anderson *et al.* (2011) and Baselga & Orme (2012). Different equations have been proposed to measure that variation. Vellend (2001) and Anderson *et al.* (2011) pointed out that studies of beta diversity might focus on two aspects of community structure, distinguishing two types of beta diversity. The first is turnover, or the directional change in community composition from one sampling unit to another along a predefined spatial, temporal, or environmental gradient. The second is variation in community composition among sampling units, which is a non-directional approach because it does not make reference to any explicit gradient. Both approaches are legitimate.

Regardless of whether beta diversity is defined as directional or non-directional, one can be interested in summarizing it using a single number that quantifies the variation. A lot of interest has been centred on the choice of the best index to produce that number. In the directional approach, the slope of the similarity decay in species composition with geographical distance can be used as a measure of beta (Nekola & White 1999). In his 1960 paper, Whittaker suggested to compute a non-directional beta index for species richness as $\beta = \gamma/\alpha$ where γ is the number of species in the region and α is the mean number of species at the study sites within the region. Since then, several other indices have been suggested to estimate a value corresponding to beta in the turnover and non-directional frameworks; see Vellend (2001), Koleff *et al.* (2003) and Anderson *et al*. (2011) for reviews. Currently, the most popular indices belong to two families that can be labelled the additive ($H_\alpha + H_\beta = H_\gamma$) and multiplicative ($H_\alpha \times H_\beta = H_\gamma$) approaches (Jost 2007, Chao *et al*. 2012). A detailed discussion of these two families is found in a *Forum* section published by *Ecology* (2010:1962–1992).

In his introduction to the *Forum*, Ellison (2010) noted that in the additive and multiplicative approaches, beta is a derived quantity that is numerically related to alpha and gamma. He pointed out that it would be most useful to have a method to estimate beta diversity without prior computation of alpha and gamma; he called for computational independence, which does not imply statistical independence. The approach adopted and developed in the present paper is to use the total variance of the site-by-species community table **Y** as a single-number estimate of beta diversity (Pelissier *et al.* 2003, Legendre *et al.* 2005, Anderson *et al.* 2006). Fulfilling Ellison's wish, it is computed without reference to the values of alpha and gamma and its statistical dependence on gamma can be accounted for using null models (Kraft *et al*. 2011, De Cáceres *et al*. 2012). While acknowledging that other measures of beta can also achieve computational and

90 statistical independence (e.g., Chao *et al*. 2012), one of our aims is to stress an important
91 advantage of the total variance of **Y** over other measures: it allows ecologists to go beyond the
92 single-number approach and partition the spatial variation in several ways to answer precise
93 ecological questions and test hypotheses about the origin and maintenance of beta diversity in
94 ecosystems.

95 We will explore the advantages and limitations of estimating beta diversity (BD_{Total}) as the
96 total variation of the community matrix **Y**. (1) In a first section, we show that BD_{Total} can be
97 obtained in two equivalent ways, i.e. by computing the sum of squares of the species occurrence
98 or abundance data or *via* a dissimilarity matrix. When the first method is used, species
99 abundances should be transformed in an appropriate way before computing BD_{Total}. The second
100 method is also appealing because it allows the estimation of beta using the dissimilarity functions
101 that are appropriate for the analysis of community data. (2) There are, however, many different
102 dissimilarity coefficients, and not all of them are appropriate for estimating beta diversity. A
103 comparative analysis of 16 coefficients is undertaken in the next section to guide users faced with
104 the problem of choosing a coefficient. (3) We then present an example to illustrate the calculation
105 of beta as the total variance of **Y** and the contributions of individual species and sampling units.
106 (4) Following that, we show that the proposals of Whittaker (1972) and Ricotta & Marignani
107 (2007) are special cases of BD_{Total} computed from a dissimilarity matrix, and that the beta
108 diversity statistic of Anderson *et al.* (2006) is closely related to BD_{Total}. (5) Finally, we show that
109 the total variance of **Y** links beta diversity assessment with the description (through ordination)
110 and hypothesis testing (through regression and canonical analysis) phases of community ecology,
111 as well as other variance partitioning methods.

112 **BETA DIVERSITY AS THE TOTAL COMMUNITY COMPOSITION VARIANCE**

113 ## Equivalent ways of computing Var(Y)

114 This section presents two equivalent ways of computing the total variance of the community
115 composition matrix **Y**. The first one is straightforward, it is simply the total variance of matrix **Y**.
116 The second one is based upon community dissimilarity matrices computed using the indices
117 developed by ecologists over more than a century. The section also shows that the total variance
118 can be divided into the contributions of individual species and individual sampling sites. Readers
119 can follow the explanation on the diagram in Fig. 1.

120 Let $\mathbf{Y} = [y_{ij}]$ be a data table containing the presence-absence or the abundance values of p
121 species (column vectors $\mathbf{y}_1$, $\mathbf{y}_2$, … $\mathbf{y}_p$ of **Y**) observed in n sampling units (row vectors $\mathbf{x}_1$, $\mathbf{x}_2$, … $\mathbf{x}_n$
122 of **Y**). We will use indices i and h for sampling units, index j for species, and y_{ij} for individual
123 values in **Y**. The total variance of **Y**, noted Var(**Y**), can be computed as follows:

124 *1. Sums of squares.* — The usual way to obtain Var(**Y**) consists in computing a matrix of
125 squared deviations from the column means. Let **S** (for "square") be a $n \times p$ rectangular matrix
126 where each element s_{ij} is the square of the difference between the y_{ij} value and the mean value of
127 the corresponding jth species:

128 $$s_{ij} = \left(y_{ij} - \bar{y}_j \right)^2 . \qquad (1)$$

129 All s_{ij} values in column j are zero if all sites have the same abundance for species j. If we sum all
130 values of **S**, we obtain the total sum of squares (SS) of the species composition data:

131 $$\mathrm{SS}_{\mathrm{Total}} = \sum_{i=1}^{n} \sum_{j=1}^{p} s_{ij} . \qquad (2)$$

132 This quantity forms the basis of BD_{Total}, which is the index of beta diversity whose properties
133 are studied in this paper:

134 $$BD_{Total} = Var(\mathbf{Y}) = SS_{Total}/(n-1) \ . \qquad (3)$$

135 Equation 3 converts the sum of squares into the usual unbiased estimator of the variance, whose
136 values can be compared between data matrices having different numbers of sampling units.
137 SS_{Total} and $Var(\mathbf{Y}) = BD_{Total}$ were both proposed by Legendre *et al.* (2005) as measures of beta
138 diversity. The two indices are equally useful to compare repeated surveys of a region involving
139 the same sites, or for simulation studies, but there is a clear advantage in using Var(**Y**) for
140 comparisons among regions.

141 Although we advocate using Var(**Y**) as a measure of beta diversity, it is important to note
142 that eqs 1-3 should not be computed directly on raw species abundance or biomass data. Because
143 calculating Var(**Y**) on raw species abundances entails that the dissimilarity between sites is
144 assessed using the Euclidean distance (eq. 7) and this coefficient is not appropriate for
145 compositional data (see section "Dissimilarity coefficients and beta assessment"), species
146 abundance data should be transformed in an ecologically meaningful way before BD_{Total} is
147 calculated using eqs 1-3.

148 An advantage of conceiving beta as the total variation in **Y** is that SS_{Total} allows the
149 assessment of the *contributions of individual species* and of *individual sampling units to the*
150 *overall beta diversity*. That is, one can compute the sum of squares corresponding to the *j*th
151 species,

152 $$SS_j = \sum_{i=1}^{n} s_{ij} \qquad (4a)$$

which is the contribution of species *j* to the overall beta diversity. SS_j divided by $(n-1)$ is the variance of species *j*. The *relative* contribution of species *j* to beta, which we call *Species Contribution to Beta Diversity* (SCBD), is thus:

$$SCBD_j = SS_j / SS_{Total} \quad . \qquad (4b)$$

In an analogous way, one can compute the sum of squares corresponding to the *i*th sampling unit,

$$SS_i = \sum_{j=1}^{p} s_{ij} \quad . \qquad (5a)$$

The SS_i values represent a genuine partitioning of beta diversity among the sites. Because the s_{ij} values are squared deviations from the species means, SS_i is the squared distance of sampling unit *i* to the centroid of the distribution of sites in species space. SS_i also measures the leverage of site *i* in a PCA ordination. The *relative* contribution of sampling unit *i* to beta diversity, which we call *Local Contribution to Beta Diversity* ($LCBD_i$), is thus:

$$LCBD_i = SS_i / SS_{Total} \quad . \qquad (5b)$$

LCBD values can be mapped, as will be shown in the ecological illustration below. Ecologically, they represent *the degree of uniqueness of the sampling units in terms of community composition*. Mapping the centred values using different symbols or colours is a way to highlight the sites with LCBD values higher and lower than the mean.

LCBD indices can be tested for significance by random, independent permutations within the columns of matrix **Y**; testing the $LCBD_i$ is the same as testing the SS_i indices. This permutation method tests H_0 that the species are distributed at random, independently of one another, among the sites, while preserving the species abundance distributions found in the observed data. However, it destroys the association of the species to the site ecological

conditions, as well as the spatial structure of community composition resulting from assembly processes (e.g. dispersal, environmental filtering, etc.). Note that the species richness (alpha diversity) of the sites is changed by this permutation method; species-poor sites become richer in most permutations and species-rich sites become poorer. Arguably, these two kinds of sites may have large LCBD for that reason, so this permutation method includes randomization of species richness in its null hypothesis. Other null hypotheses may be tested using other permutation schemes, for example by preserving site attributes such as total species richness or number of individuals (e.g. in De Cáceres *et al*. 2012). A simulation study that we performed showed that the LCBD test described here has correct rates of type I error for all coefficients that are suitable for beta diversity study (identified in section "Comparative study").

Hence, the two decompositions of SS_{Total} are:

$$SS_{Total} = \sum_{j=1}^{p} SS_j \quad \text{and} \quad SS_{Total} = \sum_{i=1}^{n} SS_i \ . \qquad (6a,b)$$

2. Dissimilarity. — As mentioned above, there is an alternative path starting from **Y** and leading to SS_{Total} (Fig. 1). That is, SS_{Total} can also be obtained from an $n \times n$ symmetric dissimilarity matrix $\mathbf{D} = [D_{hi}]$ containing Euclidean distances among points, computed using the classical Euclidean distance formula:

$$D_{hi} = D(\mathbf{x}_h, \mathbf{x}_i) = \sqrt{\sum_{j=1}^{p} (y_{hj} - y_{ij})^2} \ . \qquad (7)$$

The following equivalence is described in Legendre *et al.* (2005) and in Legendre & Legendre (2012, Chapter 8):

$$SS_{Total} = \frac{1}{n} \sum_{h=1}^{n-1} \sum_{i=h+1}^{n} D_{hi}^2 \ . \qquad (8)$$

194 That is, one can obtain SS_{Total} by summing the squared distances in the upper or lower half of
195 matrix **D** and dividing by the number of objects *n* (*not* by the number of distances). This equality
196 (eq. 8) is demonstrated in Appendix 1 of Legendre & Fortin (2010).

197 The Euclidean distance has long been known to be inappropriate for the analysis of
198 community composition data (see next section). For that reason, equations 7-8 should not be used
199 to compute SS_{Total} unless species abundance data have been appropriately transformed so that the
200 resulting dissimilarity assessments are ecologically meaningful (e.g. using the Hellinger or chord
201 transformations described in Appendix S1 in *Supporting Information*). Equation 8 can also be
202 generalized to distance matrices obtained using other dissimilarity indices. These indices may or
203 may not have the Euclidean property (P13 below), but their other properties may make them
204 appropriate for beta diversity assessment. Thus, a valid method to calculate BD_{Total} consists in
205 computing a dissimilarity matrix **D** using a selected ecological dissimilarity coefficient instead of
206 the Euclidean distance, and applying eq. 8 to obtain SS_{Total}, followed by eq. 3. That eq. 8 applies
207 to ecological dissimilarities that have the Euclidean property, or not, is shown in Appendix S2.
208 How to choose an appropriate dissimilarity coefficient for a given study is described in the next
209 section.

210 It is possible to calculate the contributions of individual sampling units from **D**. Indeed, the
211 algebra of principal coordinate analysis (PCoA, Gower 1966) offers a way of computing the sum
212 of squares SS_i, corresponding to each sampling unit *i*, directly from **D**. In PCoA, prior to eigen-
213 decomposition, the distance matrix is transformed into matrix $\mathbf{A} = [a_{hi}] = [-0.5 D_{hi}^2]$, then centred
214 as proposed by Gower (1966) using the equation

215 $$\mathbf{G} = \left(\mathbf{I} - \frac{\mathbf{1}\mathbf{1}'}{n}\right) \mathbf{A} \left(\mathbf{I} - \frac{\mathbf{1}\mathbf{1}'}{n}\right) \qquad (9)$$

where **I** is an identity matrix of size n, **1** is a vector of ones (length n) and **1**' is its transpose (Legendre & Legendre 2012, eqs 9.40 and 9.42). The diagonal elements of matrix **G** are the SS_i values, or the squared distances of the points to the multivariate centroid of **Y**, which is located at the centroid of the principal coordinate space:

$$\left[SS_i\right] = \mathrm{diag}(\mathbf{G}) \; . \qquad (10a)$$

The vector of local contributions of the sites to beta diversity ($LCBD_i$) is computed as follows:

$$\left[LCBD_i\right] = \mathrm{diag}(\mathbf{G}) / SS_{\mathrm{Total}} \; . \qquad (10b)$$

Despite its advantages, working from matrix **D** instead of the matrix of squared centred values **S** entails the drawback that one looses track of the species. Because **D** is computed among sampling units over all species, the contributions of individual species cannot be recovered from **D**.

To summarize:

• The community data table **Y** should be transformed in an appropriate way before beta diversity is computed. One can then compute the total sum of squares in the community data **Y**, SS_{Total}, from either the transformed community composition matrix **Y** (eqs. 1 and 2) or from a Euclidean distance matrix **D** computed from the transformed data (eqs. 7 and 8). The two modes of calculation produce the same statistic, SS_{Total}, and from it one can compute the total variance, $BD_{\mathrm{Total}} = \mathrm{Var}(\mathbf{Y})$ (eq. 3).

• Alternatively, one can use eq. 8 to compute SS_{Total} from a dissimilarity matrix **D** obtained using any appropriate dissimilarity coefficient (next section). Equation 8 applies to ecological dissimilarity indices that have the Euclidean property, or not, as demonstrated in Appendix S2.

237 • The contribution of the *i*th sampling unit to the overall beta diversity can be computed using
238 eq. 5a. From these, *Local Contribution to Beta Diversity* (LCBD) coefficients can be derived.
239 LCDB coefficients are comparative indicators of the ecological uniqueness of the sites in terms
240 of community composition. The SS_i values are also found on the diagonal of matrix **G** (eqs 9 and
241 10a). The relative contributions (LCDB) are computed using eqs 5b and 10b.

242 • If BD_{Total} is calculated from **Y** (eq. 3) transformed in an appropriate way, the contribution of
243 species *j* to the overall beta diversity, SS_j, is computed using eq. 4a, and the relative
244 contributions, called the *Species Contributions to Beta Diversity* (SCBD), are computed using eq.
245 4b. SCBD coefficients represent the degree of variation of individual species across the study
246 area. SS_j and SCDB coefficients cannot be derived from a distance matrix.

247 **DISSIMILARITY COEFFICIENTS AND BETA ASSESSMENT**

248 Since the description of the first floristic similarity coefficient by Paul Jaccard (1900),
249 community ecologists have developed a broad array of similarity and dissimilarity coefficients.
250 Ecologists are often faced with the question: Which community data transformation and/or
251 (dis)similarity coefficient should I use in my study? When assessing beta diversity through the
252 variation in community composition, one needs to specify what is meant by "variation in
253 community composition". The answer will determine the choice of a community data
254 transformation and/or dissimilarity measure, and must be carefully articulated (Anderson *et al.*
255 2006).

256 There is no single coefficient that is appropriate in all occasions. Choice should be guided
257 by the properties of coefficients and the objective of the research. Several studies have compared
258 resemblance coefficients, focussing on their linearity and resolution along simulated gradients

(e.g. Bloom 1981, Hajdu 1981, Gower & Legendre 1986, Faith *et al.* 1987, Legendre & Gallagher 2001), or investigating theoretical properties (e.g. Janson & Vegelius 1981, Hubálek 1982, Wilson & Shmida 1984, Gower & Legendre 1986, Koleff *et al.* 2003, Chao *et al.* 2006, Clarke *et al.* 2006). Complementing these studies, we present in this section a comparative review of several abundance- and incidence-based dissimilarity coefficients, listed in Table 1. Our aim is to determine which coefficients are the most appropriate for assessing beta diversity under the present approach. We restricted the list to the coefficients originally designed for pairwise comparisons, thus excluding multiple-site dissimilarity measures (e.g. Baselga 2010, 2013). In addition, we focused on properties that are easy to understand and interpret ecologically, with preference for those that could be checked unequivocally.

Properties of dissimilarity indices for the study of beta diversity

Fourteen properties, divided in four groups, are described in Appendix S3, which also outlines procedures to check which dissimilarity indices possess them. The first two groups (P1 to P9) contain the minimum requirements for assessing beta diversity. The remaining two groups (P10 to P14) are not necessarily required in all beta diversity studies. Practitioners should determine whether the context of their analyses requires these properties, or not. Other properties are also considered interesting by authors of other studies on dissimilarity coefficients.

The dissimilarity coefficients

A selection of 16 quantitative dissimilarity coefficients commonly used for beta diversity assessment was considered in our comparison study. They represent a broad hand among the available coefficients. Equations are shown in Table 1 for community composition abundance and for presence-absence (i.e. incidence) data. Table 2 indicates which dissimilarity coefficients

possess the properties mentioned in the previous paragraph and described in Appendix S3, as well as their maximum values (D_{max}) when they exist.

The first coefficient in the list is the Euclidean distance. Although this distance is known to be inappropriate for the analysis of community composition data sampled under varying environmental conditions (Orlóci 1978, Legendre & Gallagher 2001), it is included in the comparison where it will serve as a reference point. It is the failure of the Euclidean distance to correctly account for beta diversity (it lacks properties P4, P5, P7, P8, P9) that makes it necessary for ecologists to rely on the other dissimilarity measures investigated in this paper. The Euclidean distance may, however, become appropriate after transformation of the community data (Appendix S1). Likewise, the Manhattan distance is inappropriate *per se*; nevertheless, it is included in the comparison because it becomes the Whittaker dissimilarity after profile transformation of **Y**, and that index is appropriate for beta diversity studies (Whittaker 1952, Faith *et al.* 1987; Appendix S1).

The other coefficients included in the comparative study are double-zero asymmetric (property P4); they have been recommended and used for community composition assessment or beta diversity studies. Four of these dissimilarities can be computed using the formula in Table 1 or through the alternative method corresponding to property P14. For the species profile, Hellinger, chord, and chi-square distances, the data are first transformed using the same-name transformation (Appendix S1); computing the Euclidean distance (eq. 7) on the transformed data produces the targeted profile, Hellinger, chord, or chi-square distance.

When applied to presence-absence data, several quantitative coefficients in Table 1 produce either the one-complement of the Jaccard similarity index or the one-complement of the Sørensen index. The Hellinger and chord distances both produce $D = \sqrt{2(1 - \text{Ochiai similarity})}$.

Comparative study

The properties of the selected coefficients were coded into a data matrix with the coefficients as rows and properties P4 to P14 as columns (Table 2). Most properties were coded as presence-absence (0-1), except for P13 which was coded on a semiquantitative 0-1-2 scale (0 = not Euclidean, 1 = $\mathbf{D}^{(0.5)}$ is Euclidean, 2 = $\mathbf{D}^{(0.5)}$ and **D** are Euclidean). The missing value in Table 2 (coded "NA") was transformed to 1; the reason is that the chi-square distance has property P7, so it would likely have P10 if a binary form was available for that coefficient. The data matrix was subjected to principal component analysis (PCA) of the correlation matrix.

The analysis produced an ordination of the dissimilarities (Fig. 2) where similar coefficients are close to one another and dissimilar ones are more distant. Properties P4 to P14, which are the variables of the matrix subjected to PCA, are shown as red arrows. One can identify five types of coefficients using the data in Table 2 and the ordination diagram:

• Type I contains the Euclidean and Manhattan distances, as well as the mean character difference and the species profile distance. They all lack several of the important properties in the first two classes (P4 to P9). Most notably, the Euclidean and Manhattan distances do not have the double-zero asymmetry property (P4), and the four coefficients fail to give the largest dissimilarity values to pairs of sites without species in common (P5). The distance between species profiles decreases when the number of unique species in the compared sites increases (P6). The Euclidean distance, Manhattan distance and species profile distances are not species-replication invariant (P7). Moreover, the Euclidean, Manhattan, and modified mean character difference do not fulfil P8 and P9. The species profile distance is invariant to the measurement units of the data (P8), but the upper bound of $\sqrt{2}$ is only reached when there is a single, unique species per site; with more

species, the maximum distance decreases with the number of unique species. Due to these shortcomings, the four coefficients belonging to type I do not allow proper assessment and comparison of beta diversity estimates among data sets.

Coefficients in types II to IV provide asymmetrical treatment of double-zeros (P4) and they all have properties P5 to P9, which are required for comparability of beta estimates among data sets. They are thus all appropriate for beta diversity assessment.

• Type II contains the Hellinger and chord distances. These two distances are closely related: the Hellinger distance is equal to the chord distance computed on square-root-transformed species frequencies. They share all properties in classes 1 and 2, which are necessary for beta diversity assessment. Furthermore, type II coefficients are Euclidean (P13) and they can be emulated by transformations of the raw frequency or biomass data (P14). Hence, **D** matrices computed using these coefficients are fully suitable for ordination by principal coordinate analysis (PCoA), which will not produce negative eigenvalues and complex axes. For an easier and more informative ordination, species frequency (or frequency-like, such as biomass) data transformed using the Hellinger and chord transformations (Appendix S1) can be analysed directly by principal component analysis (PCA) and by canonical redundancy analysis (RDA); this is not the case for the type III and IV coefficients. (PCoA of Hellinger and chord distance matrices produces the same ordinations as PCA of the Hellinger and chord transformed data.) Moreover, SS_{Total} corresponding to the Hellinger and chord distances can be obtained by computing the transformation in Appendix S1, then applying eqs 1 and 2 to the transformed data. This is simpler than computing the distance matrix and using eq. 8 to obtain SS_{Total}. Furthermore, the Hellinger and chord transformed data allow the computation of SCBD statistics (eq. 4b), which cannot be obtained from a distance matrix.

• Type III contains the divergence, Canberra, Whittaker, percentage difference (*alias* Bray-Curtis), Wishart and Kulczynski dissimilarities. They share all properties in classes 1 and 2, which are necessary for beta diversity assessment. The coefficient of divergence, which is Euclidean, can be used directly in PCoA ordination. For four coefficients (Canberra, Whittaker, percentage difference and Wishart), the square root of the distances must be taken before they are used in PCoA. The matrix of principal coordinates can be used as the response data in RDA; this is the distance-based RDA method proposed by Legendre & Anderson (1999). Among the six coefficients in this group, only the Whittaker index is invariant to the total abundance of each sampling unit (P11); the remaining indices are thus affected to some extent by differences in total abundances between the two compared sites. The Kulczynski coefficient is suitable for beta diversity assessment, but not for ordination, and it does not correct for undersampling. Considering the properties analysed in this paper, this coefficient does not offer any particular advantage not available in other coefficients, and its application is limited.

• Type IV contains the abundance-based quantitative forms of the Jaccard, Sørensen and Ochiai indices. Like coefficients of type II, type IV coefficients fulfil property P11 (invariance to total abundance in individual sampling unit). In addition, they have property P12 (correction for undersampling), but not properties P13 and P14, which are desirable for ordination. In particular, type IV coefficients are not Euclidean (P13) in quantitative form, although the Jaccard, Sørensen and Ochiai similarities, which are the binary counterparts of the first three, produce coefficients with the Euclidean property when transformed to $D = \sqrt{1 - \text{similarity}}$ (Legendre & Legendre 2012, Table 7.2).

• The chi-square distance forms type V. This distance is widely used to analyse communities since it is the basis for correspondence analysis. The chi-square distance gives more importance

372 to rare than common species in the assessment of the distance between sites, the rare species
373 (when their abundances are correctly estimated by sampling) being considered as more important
374 indicators of special environmental conditions prevailing at some sites. Unfortunately, it lacks
375 property P5, and this makes it unsuitable for beta diversity studies.

376 Maximum value of BD

377 All dissimilarities in types II to IV have a maximum value, reached when two sites have
378 completely different community compositions. For example, the Hellinger and chord distances in
379 type II have a minimum value of 0 and a maximum of $\sqrt{2}$ (Table 2). If all sites have entirely
380 different species compositions, all $n(n-1)/2$ distances in **D** are $\sqrt{2}$ and eqs 8 and 3 produce
381 $BD_{Total} = 1$. Hence for these two dissimilarity indices, BD_{Total} is in the range [0, 1]. All other
382 indices that are appropriate for beta assessment (types III and IV) have maximum values of 1.
383 When all sites have different species compositions, the distances are all equal to 1 and BD_{Total}
384 computed through eqs 8 and 3 is 0.5, so that BD_{Total} is in the range [0, 0.5]. For these distances,
385 multiplying BD_{Total} by 2 would directly produce relative BD values (BD_{rel}, Appendix S3,
386 property P9) in the range [0, 1]. Hence BD_{Total} has a fixed range of values for any community,
387 which does not depend on the total abundance in the community composition table.

388 ECOLOGICAL ILLUSTRATION: FISH BETA DIVERSITY IN DOUBS RIVER

389 Freshwater fish were collected by Verneaux (1973) in the Doubs River, a tributary of the Saône
390 that runs near the France-Switzerland border in the Jura Mountains in eastern France. In his
391 paper, Verneaux proposed to use fish communities to characterize ecological zones along
392 European rivers and streams. The data include fish community composition at 30 sites along the
393 453 km course of the river, the site geographic coordinates, and environmental data (source:

394 http://www.bio.umontreal.ca/numecolR/). 27 species were captured and identified. No fish
395 were caught at site 8, hence that site was excluded from the reanalyses made by Borcard *et al.*
396 (2011), as well as here. As in that book, we subjected the fish data to a chord transformation
397 before analysis (Appendix S1).

398 SS_{Total} (eq. 2) was 15.243 and BD_{Total} (eq. 3) was 0.544 for the fish data. The local
399 contributions of individual sites were computed; the values of SS_i (eq. 5a) ranged from 0.291 to
400 0.971. An ordination diagram (Fig. 3) illustrates the mathematical meaning of SS_i indices: they
401 are the squares of the distances of the sites to the multivariate centroid, as discussed under eq. 9.

402 The relative contributions ($LCBD_i = SS_i / SS_{Total}$, eq. 5b) were in the range [0.019, 0.064].
403 LCBD indices indicate the uniqueness of the fish community at each site. They are plotted on a
404 schematic map of the river in Fig. 4a, which also shows the two sites where LCBD was
405 statistically significant. Comparison with species richness (Fig. 4b) showed that for this data,
406 LCBD was negatively correlated to richness ($r = -0.60$), indicating that high LCBD (i.e. high
407 uniqueness of species composition) was often related to a small number of species. This is not,
408 however, a general or obligatory relationship.

409 Environmental variables were also available for each site: distance from the source,
410 altitude, riverbed slope, mean minimum discharge, pH, concentrations in calcium, phosphate,
411 nitrate, ammonium and dissolved oxygen, and biochemical oxygen demand (BOD). The LCBD
412 values were regressed on the environmental variables to determine the factors that make LCBD
413 vary along the river (adjusted $R^2 = 0.58$). Only two environmental variables were retained by
414 backward elimination in regression: riverbed slope and BOD. Both variables had positive
415 coefficients in the model, indicating that sites with high BD_{Total} either had a large slope (specially
416 true at the headwaters) or were strongly eutrophic (high BOD). Note that regressing LCBD

values on environmental variables is not the same as canonical analysis of the community data. For the chord-transformed Doubs fish data, forward selection of environmental variables in RDA produced a different model (adjusted $R^2 = 0.61$) containing five significant variables at the 0.05 level: distance from the source, altitude, slope, dissolved oxygen, and BOD. The question in RDA is to identify the factors driving the observed variation in community composition; RDA truly analyses beta diversity by decomposing the total variance of the species data, i.e. BD_{Total}, into explained and residual components. By contrast, in regression analysis of the LCBD indices, the question is why some sites have higher degrees of uniqueness in species composition than others.

Four species contributed to beta diversity well above the mean of the 27 species: the stone loach (*Barbatula barbatula*, Balitoridae), the common bleak (*Alburnus alburnus*, Cyprinidae), the Eurasian minnow (*Phoxinus phoxinus*, Cyprinidae), and the brown trout (*Salmo trutta fario*, Salmonidae) which had the highest SCBD index. The chord-transformed abundances of these species varied the most among sites. The brown trout, Eurasian minnow and stone loach are found in the unpolluted sites with high LCBD upriver, which have high conservation status, whereas the common bleak is abundant in the eutrophic sites with agricultural pollution in the middle course of the river. Sites in the latter group, which also have high LCBD values, are in need of restoration.

One may wonder: For the coefficients that are appropriate for beta diversity studies, are the LCBD estimates similar or very different? Using the software in Appendix S4, calculation of LCBD was repeated for the 11 dissimilarities belonging to types II to IV, which are appropriate for beta assessment. The 11 LCBD vectors were quite similar: their mean Spearman correlation was 0.905. Kendall concordance analysis (Legendre 2005) showed that the contributions of all 11

vectors to the concordance of the group were significant. (These are not genuine tests of significance since the LCBD vectors were all computed from the same data; these results provide, however, a clustering validation criterion.) These results show that LCDB indices computed using all dissimilarities that were suitable for beta diversity assessment were highly concordant.

DISCUSSION

Different concepts of beta diversity

We will first address the appropriateness of using "beta diversity" to designate the approach described in this paper. We acknowledge that this is an unsettled issue. Authors, e.g. Anderson *et al.* (2011), have rightfully argued that there are several meanings and measures associated with the concept of beta diversity. Authors agree that alpha and beta diversities are essentially different; alpha measures how diversified the species are within a site, i.e. in a single row of the site-by-species data table **Y**, whereas beta measures how diversified the sites are in species composition within a region, i.e. the variation among the rows of **Y**. Some ecologists prefer to reserve the expression beta diversity for the additive or multiplicative approaches, and we will not dispute their choice.

However, if beta diversity can be seen as "the variation in species composition among sites", as stated by many authors, then the variance of **Y**, which specifically measures that variation, certainly qualifies as a measure of beta. The literature is growing that adopts this broader concept and measure of beta, because it links the ecological concept of beta diversity to methods of analysis that can be applied to test hypotheses about the mechanisms that generate and maintain beta diversity in ecosystems (subsection "Multiple ways of partitioning total beta diversity"). Those who prefer to limit the meaning of beta diversity to the additive or

462 multiplicative approaches do not deny that variation in species composition among sites can be
463 analysed, and hypotheses tested, but they prefer to call that variation by some other name, e.g.
464 compositional heterogeneity among sites. Compositional heterogeneity — be it called beta
465 diversity, or not — measures community differentiation, which results from evolutionary and
466 ecological processes operating at several spatial (from site to global) and temporal scales.

467 After proposing the concept in his seminal papers, Whittaker (1960, 1072) detailed
468 different measures of beta diversity. One of his measures corresponds precisely to the variance of
469 **Y** measured through some dissimilarity coefficients, as will be shown in the next subsection. We
470 are in good company here. Ecologists largely agree with Whittaker (1972) that beta diversity
471 conceptually corresponds to *the variation in species composition among sites in the geographic*
472 *region of interest*. [Whittaker used a slightly different expression, "the extent of differentiation of
473 communities along habitat gradients". He was interested in the response of communities to
474 environmental variation, hence his interest for ordination methods.] Legendre *et al.* (2005) were
475 perhaps the first to use precisely that expression, based on their reading of Whittaker, and they
476 were followed in its use by many authors, including Anderson *et al.* (2006) and Anderson *et al.*
477 (2011). Leaving the terminological issue aside, we may discuss what are the different ways of
478 estimating the variation in species composition among sites, or beta diversity. For example,
479 Baselga (2013) suggested calculating multiple-site dissimilarity coefficients to measure variation
480 in species composition between more than two sites, instead of using an average of pairwise
481 dissimilarity values. Alternative estimation methods are not in opposition but complementary;
482 each one offers a different way of explaining beta diversity, or expressing it in a way that makes
483 it useful for ecological interpretation, impact assessment, or conservation studies. Future studies

484 should focus on comparing alternative estimation approaches in order to clarify their
485 differences and domains of application.

486 Related approaches to beta diversity assessment

487 In this paper, we used the total variance of **Y** as an estimate of beta diversity (BD_{Total}) for a region
488 of interest (eq. 3, Fig. 1). Var(Y) should not be computed using raw abundance data but after
489 some appropriate transformation of the community composition data, or through a carefully
490 selected dissimilarity function. The values of BD_{Total} are comparable among data sets having the
491 same or different numbers of sampling units (n), provided that the sampling units are of the same
492 size or represent the same sampling effort, and that the calculations have been done using the
493 same index chosen among those that have been found to be suitable for beta diversity assessment
494 in this paper. Depending on the index, BD_{Total} may have a maximum value of 1 or 0.5 when all
495 sites under study have different species compositions.

496 Alternative equations to estimate total BD have been proposed by Whittaker (1972),
497 Ricotta & Marignani (2007) and Anderson *et al.* (2006). We will now show that these proposals
498 are special cases of eq. 3 or are related to it.

499 In section "Equivalent ways of computing Var(**Y**)", we saw that SS(**Y**) can be computed as
500 the sum of the squared dissimilarities divided by n (eq. 8). This is appropriate for the Euclidean
501 distance and for dissimilarities that have the property of being Euclidean (P13). Appendix S2
502 shows that SS_{Total} can also be computed in that way for dissimilarities that do not lead to a fully
503 Euclidean representation; these will not concern us in the present paragraph. Several
504 dissimilarities, coded 1 for P13 in Table 2, are Euclidean only when taking their square roots; the
505 square-rooted distances form matrix $\mathbf{D}^{(0.5)} = \left[D_{hi}^{0.5}\right]$. That group includes the Canberra metric,

506 Whittaker's index, the percentage difference (*alias* Bray-Curtis) and Wishart's coefficient.
507 Many of the incidence-based (i.e. binary) coefficients are also in that situation, including the
508 widely used Jaccard, Sørensen and Ochiai coefficients (Legendre & Legendre 2012, Table 7.2).
509 We will show here that the method of calculation of beta diversity proposed in other papers is
510 equivalent to DB_{Total} of the present paper if $\mathbf{D}^{(0.5)}$ is used for the calculation.

511 (a) Whittaker (1972, p. 233) stated that "The mean CC [Jaccard or Sørensen coefficient of
512 community] for samples of a set compared with one another in all possible directions is one
513 expression [of] their relative dissimilarity, or beta differentiation". The mean is obtained by
514 summing the dissimilarities and dividing by the number of dissimilarities in the half-matrix,
515 $n(n-1)/2$. This is equivalent to computing eqs 8 and 3 on the square-rooted dissimilarities
516 (matrix $\mathbf{D}^{(0.5)}$) and multiplying by 2. Hence, Whittaker's formula only differs by a factor 2 from
517 DB_{Total} computed from $\mathbf{D}^{(0.5)}$.

518 (b) There is also a relationship between the equation for DB_{Total} used in this paper and the
519 suggestion of Ricotta & Marignani (2007) to estimate beta diversity by Rao's (1982) quadratic
520 entropy, $Q = \sum_{h=1}^{n-1} \sum_{i=h+1}^{n} \delta_{hi} p_h p_i$, where p_i and p_h contain the relative abundance of sampling
521 units *i* and *h*, respectively, and δ_{hi} is the dissimilarity between *i* and *h* computed with any measure
522 of one's choice. If all sampling units are considered to be equally important, say $p_i = 1/n$,
523 then $Q = \frac{1}{n^2} \sum_{h=1}^{n-1} \sum_{i=h+1}^{n} \delta_{hi}$, which is very close to DB_{Total} computed from $\mathbf{D}^{(0.5)}$ through eq. 8
524 followed by eq. 3. The difference is that the last division is by *n* in *Q* instead of $(n-1)$ in eq. 3.

525 (c) The beta diversity statistic developed by Anderson *et al.* (2006) belongs to the same family as
526 DB_{Total}. It is the *sum of the dissimilarities from the sampling units to the group centroid* in

multivariate space divided by *n*, producing a maximum likelihood estimate of the variance. It differs from DB_{Total}, which is the *sum of the squared dissimilarities from the sampling units to the group centroid* divided by $(n-1)$ (eq. 3). The squared dissimilarities from the sampling units to the group centroid are found in vector $[SS_i]$ obtained by eqs 9 and 10a computed from **D**. Because it can be computed from any dissimilarity matrix, the Anderson *et al.* (2006) statistic can be computed from **D** or $\mathbf{D}^{(0.5)}$, both producing a different statistic than DB_{Total}.

Regarding the choice of a dissimilarity measure and the equivalence of the beta diversity approaches described in the last paragraphs, different situations should be considered. (a) For dissimilarity measures that are not Euclidean for **D** but are Euclidean for $\mathbf{D}^{(0.5)}$, then the approaches of Whittaker (1972) and Ricotta & Marignani (2007) are essentially equivalent to the calculation of DB_{Total} in the present paper. (b) If the dissimilarity measure can be obtained by applying a transformation to the original data (Appendix S1) followed by the computation of the Euclidean distance, the equivalence between these methods holds in the transformed space and BD_{Total} can be computed by applying eqs 2 and 3 to the transformed data. (c) If the dissimilarity measure cannot be obtained by applying a transformation to the original data followed by Euclidean distance calculation, the distances to the centroid can still be computed using the square root of eq. 10a. This result holds for non-Euclidean embeddable dissimilarities as well, although with some additional complexities (Anderson 2006; Appendix S2).

Multiple ways of partitioning total beta diversity

The strongest advantage of adopting the present approach to the analysis of beta diversity lies in the possibility of partitioning the total sum-of-squares of the community composition data into additive components. The total variance is the basic currency of many statistical methods,

549 univariate and multivariate, through which Var(**Y**) can be partitioned in different ways.
550 Available partitioning methods include the following.

551 1. *Contributions of individual species*. — The SS_{Total} statistic can be partitioned into species
552 contributions to beta diversity ($SCBD_j$, eq. 4b). SCBD indices can, in principle, be computed for
553 raw or transformed abundance data, but it should in practice be limited to data subjected to the
554 Hellinger or chord transformations, which are the only two that correspond to distances suitable
555 for beta assessment. After centring, the SCBD values have signs which indicate the species that
556 vary more [or less] than the mean across the sites. A mathematical limitation restrains the use of
557 SCBD coefficients: they can only be computed from raw or transformed data tables with species
558 in columns; they cannot be computed from a **D** matrix. Calculating SCBD indices is useful to
559 determine which species exhibit large variations across the study area. Note that SCBD indices
560 do not have the same interpretation as indicator species for groups of sites (Dufrêne & Legendre
561 1997, De Cáceres & Legendre 2009). The sites where species with large SCBD values are
562 abundant and dominate the community will normally also have large LCBD indices, as we found
563 in our example.

564 2. *Contributions of individual sampling units*. — Likewise, the SS_{Total} statistic can be partitioned
565 into local contributions of individual sampling units to beta diversity ($LCBD_i$, eq. 5b or 10b). The
566 LCBD values, which can be mapped, indicate the sites that contribute more [or less] than the
567 mean to beta diversity. LCBD are comparative indicators of site uniqueness; hence, large LCBD
568 values indicate sites that have strongly different species compositions. For conservation biology,
569 large LCBD values may indicate sites that have unusual species combinations and high
570 conservation value, or degraded and species-poor sites in need of ecological restoration. They
571 may also correspond to special ecological conditions or result from the effect of invasive species

572 on communities. LCBD may be inversely correlated with species richness, as in our example,
573 but in other ecosystems large LCBDs may indicate rare species combinations that are worth
574 studying in more detail.

575 In data analysis, sites with high LCBD may be removed before simple or canonical
576 ordination because they may have an undue influence on the results. This may prove a useful
577 criterion to remove sites prior to ordination, instead of other criteria like low species richness.

578 3. *Within- and among-group contributions*. — Groups of sites may be known *a priori* from the
579 sampling design, or they may be obtained by clustering based on the environmental variables. For
580 these groups of sites, the total sum-of-squares of the species data can be divided by multivariate
581 analysis of variance (computed using MANOVA or canonical analysis) into within- and among-
582 group sums of squares. Alternatively, groups of sites where the species respond in the same way
583 to environmental variables can be identified by multivariate regression tree analysis.

584 4. *Simple and canonical ordination*. — The total sum-of-squares, which estimates beta diversity,
585 can be partitioned into orthogonal axes by simple ordination methods (PCA, CA, PCoA).
586 Alternatively, SS_{Total} can be partitioned by canonical analysis (RDA or CCA) into orthogonal
587 axes related to the environmental variables.

588 5. *Contributions of sets of explanatory factors*. — SS_{Total} can be partitioned as a function of
589 different sets of explanatory variables by variation partitioning (Borcard *et al.* 1992; Peres-Neto
590 *et al.* 2006). Partitioning can be done, for example, between different sets of environmental
591 variables, or between explanatory matrices representing environmental and spatial variables (e.g.
592 sets of spatial eigenfunctions), depending on the hypotheses under study. This is a major
593 approach for estimating the relative contributions of groups of explanatory variables representing

different hypotheses about the origin of beta diversity.

6. *Spatial scales*. — SS_{Total} can be partitioned as a function of spatial scales by spatial eigenfunction analysis. See Legendre & Legendre (2012, Chapter 14) for a review of these methods. These and other methods of multivariate multiscale analysis were also reviewed by Dray *et al.* (2012).

7. *Multivariate variogram and multiscale ordination*. — SS_{Total} can also be partitioned into spatial scales by multivariate variogram analysis (Wagner 2003). Furthermore, the species-environment relation, which represents a portion of SS_{Total}, can be partitioned into spatial scales by multiscale ordination; see Wagner (2003, 2004) and Legendre & Legendre (2012, Section 14.4).

Choosing a dissimilarity index for beta diversity assessment

Analysing the spatial variation in species composition necessarily implies choosing a dissimilarity coefficient, either implicitly or explicitly (Legendre *et al.* 2005, Anderson *et al.* 2006). Choosing an appropriate coefficient is crucial to ensure the interpretation of the results and allow the comparison of beta diversity estimates among regions and types of organisms.

In this paper, we studied several properties of coefficients, separating those that were purely mathematical from those that had an ecological interpretation. This conceptual separation was important to help users make choices on ecological grounds. Comparison of the 16 selected dissimilarity coefficients based on 14 ecological, statistical and mathematical properties led to a model where the coefficients were divided into five main types. Three of those types are suitable for beta diversity studies and comparison of beta diversity estimates computed from different ecological data sets. These different types of coefficients can be used to address different

questions.

Among the unsuitable coefficients are the Manhattan and Euclidean distances. As shown in this paper, these distances are appropriate for beta diversity assessments only after transformation of the raw abundance data. In the case of the Manhattan distance (L1 norm), the natural transformation is the division of each value by the total abundance, which leads to the Whittaker coefficient. In the case of the Euclidean distance (L2 norm), the natural transformation is the division of each value by the norm of the row vector, which leads to the chord distance. The Hellinger distance is the chord distance computed on square-root-transformed abundance data.

When choosing a coefficient, users should check the properties the coefficient has, and determine whether they are suitable for the objectives of the study. Further research is needed about the mathematical and ecological properties of dissimilarity coefficients and the situations where these properties are desirable or needed.

ACKNOWLEDGEMENTS

This paper is dedicated to Dr. Francesc Oliva, who fostered the interest of M. De Cáceres for dissimilarity coefficients and their use in ecology. Our thanks to Daniel Borcard who provided comments on a first draft of the manuscript, and to Anne Chao and two other anonymous referees who provided very interesting comments that helped us improve the paper. This research was supported by a NSERC grant no. 7738 to P. Legendre. M. De Cáceres was supported by research projects BIONOVEL (CGL2011-29539/BOS) and MONTES (CSD2008-00040) funded by the Spanish Ministry of Education and Science.

636 **REFERENCES**

637 1.

638 Anderson, M.J. (2006). Distance-based tests for homogeneity of multivariate dispersions.

639 *Biometrics*, 62, 245–253.

640 2.

641 Anderson, M.J., Ellingsen, K.E. & McArdle, B.H. (2006). Multivariate dispersion as a measure

642 of beta diversity. *Ecol. Lett.*, 9, 683–693.

643 3.

644 Anderson, M.J., Crist, T.O., Chase, J.M., Vellend, M., Inouye, B.D., Freestone, A.L. *et al.*

645 (2011). Navigating the multiple meanings of β diversity: a roadmap for the practicing

646 ecologist. *Ecol. Lett.*, 14, 19–28.

647 4.

648 Baselga, A. (2010). Partitioning the turnover and nestedness components of beta diversity.

649 *Global Ecol. Biogeogr.*, 19, 134–143.

650 5.

651 Baselga, A. (2013). Multiple site dissimilarity quantifies compositional heterogeneity among

652 several sites, while average pairwise dissimilarity may be misleading. *Ecography*, 36,

653 124–128.

654 6.

655 Baselga, A. & Orme, C.D.L. (2012). betapart: an R package for the study of beta diversity.

656 *Methods Ecol. Evol.*, 3: 808–812.

657 7.

658 Bloom, S.A. (1981). Similarity indices in community studies: potential pitfalls. *Mar. Ecol. Progr.*

Ser., 5, 125–128.

8.

Borcard, D., Gillet, F. & Legendre, P. (2011). *Numerical ecology with R.* Use R! series, Springer Science, New York.

9.

Borcard, D., Legendre, P. & Drapeau, P. (1992). Partialling out the spatial component of ecological variation. *Ecology*, 73, 1045–1055.

10.

Bray, R.J. & Curtis, J.T. (1957). An ordination of the upland forest communities of southern Wisconsin. *Ecol. Monogr.*, 27, 325–349.

11.

Cardoso, P., Borges, P.A.V. & Veech, J.A. (2009). Testing the performance of beta diversity measures based on incidence data: the robustness to undersampling. *Divers. Distrib.*, 15, 1081–1090.

12.

Chao, A., Chazdon, R.L., Colwell, R.K. & Shen, T.J. (2006). Abundance-based similarity indices and their estimation when there are unseen species in samples. *Biometrics* 62, 361–371.

13.

Chao, A., Chiu, C.-H. & Hsieh, T.C. (2012). Proposing a resolution to debates on diversity partitioning. *Ecology*, 93, 2037–2051.

14.

Clark, P.J. (1952). An extension of the coefficient of divergence for use with multiple characters. *Copeia*, 1952, 61–64.

15.

683 Clarke, K.R., Somerfield, P.J. & Chapman, M.G. (2006). On resemblance measures for
684 ecological studies, including taxonomic dissimilarities and a zero-adjusted Bray–Curtis
685 coefficient for denuded assemblages. *J. Exp. Mar. Biol. Ecol.*, 330, 55–80.
686 16.
687 Czekanowski, J. (1909). Zur Differentialdiagnose der Neandertalgruppe. *Korrespondenz-Blatt*
688 *deutsch. Ges. Anthropol. Ethnol. Urgesch.*, 40, 44–47.
689 17.
690 De Cáceres, M. and Legendre, P. (2009). Associations between species and groups of sites:
691 indices and statistical inference. *Ecology*, 90, 3566–3574.
692 18.
693 De Cáceres, M., Legendre, P., Valencia, R., Cao, M., Chang, L.-W., Chuyong, G. *et al.* (2012).
694 The variation of tree beta diversity across a global network of forest plots. *Global Ecol.*
695 *Biogeogr.*, 21, 1191–1202.
696 19.
697 Dray, S., Pélissier, R., Couteron, P., Fortin, M.-J., Legendre, P., Peres-Neto, P.R. *et al.* (2012).
698 Community ecology in the age of multivariate multiscale spatial analysis. *Ecol. Monogr.*,
699 82, 257–275.
700 20.
701 Dufrêne, M. & Legendre P. (1997). Species assemblages and indicator species: the need for a
702 flexible asymmetrical approach. *Ecol. Monogr.*, 67, 345–366.
703 21.
704 Ellison, A.M. (2010). Partitioning diversity. *Ecology*, 91, 1962–1963.
705 22.
706 Faith, D.P., Minchin, P.R. & Belbin, L. (1987). Compositional dissimilarity as a robust measure

of ecological distance. *Vegetatio*, 69, 57–68.

23.

Gower, J.C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, 53, 325–338.

24.

Gower, J.C. & Legendre, P. (1986). Metric and Euclidean properties of dissimilarity coefficients. *J. Classif.*, 3, 5–48.

25.

Hajdu, L.J. (1981). Graphical comparison of resemblance measures in phytosociology. *Vegetatio*, 48, 47–59.

26.

Hubálek, Z. (1982). Coefficients of association and similarity, based on binary (presence-absence) data: an evaluation. *Biol. Rev.*, 57, 669–689.

27.

Hurlbert, S.H. (1984). Pseudoreplication and the design of ecological field experiments. *Ecol. Monogr.*, 54, 187–211.

28.

Jaccard, P. (1900). Contribution au problème de l'immigration post-glaciaire de la flore alpine. *Bull Soc. Vaudoise Sci. Nat.*, 36, 87–130.

29.

Janson, S. & Vegelius, J. (1981). Measures of ecological association. *Oecologia*, 49, 371–376.

30.

Janssen, J.G.M. (1975). A simple clustering procedure for preliminary classification of very large sets of phytosociological relevés. *Vegetatio*, 30, 67–71.

731 31.

732 Jost, L. (2007). Partitioning diversity into independent alpha and beta components. *Ecology*, 88,

733 2427–2439.

734 32.

735 Koleff, P., Gaston, K.J. & Lennon, J.J. (2003). Measuring beta diversity for presence-absence

736 data. *J. Anim. Ecol.*, 72, 367–382.

737 33.

738 Kraft, N.J.B., Comita, L.S., Chase, J.M., Sanders, N.J., Swenson, N.G., Crist, T.O. *et al.* (2011).

739 Disentangling the drivers of diversity along latitudinal and elevational gradients. *Science*,

740 333, 1755–1758.

741 34.

742 Kulczynski, S. (1928). Die Pflanzenassoziationen der Pieninen. *Bull. Int. Acad. Pol. Sci. Lett. Cl.*

743 *Sci. Math. Nat. Ser. B*, Suppl. II (1927), 57–203.

744 35.

745 Lance, G.N. & Willams, W.T. (1967). Mixed-data classificatory programs. I. Agglomerative

746 systems. *Aust. Comput. J.*, 1, 15–20.

747 36.

748 Lebart, L. & Fénelon, J.P. (1971). *Statistique et informatique appliquées.* Dunod, Paris, France.

749 37.

750 Legendre, P. (2005). Species associations: the Kendall coefficient of concordance revisited. *J.*

751 *Agr. Biol. Envir. S.*, 10, 226–245.

752 38.

753 Legendre, P. & Anderson, M.J. (1999). Distance-based redundancy analysis: testing multispecies

754 responses in multifactorial ecological experiments. *Ecol. Monogr.*, 69, 1–24.

39.

Legendre, P., Borcard, D. & Peres-Neto, P.R. (2005). Analyzing beta diversity: partitioning the spatial variation of community composition data. *Ecol. Monogr.*, 75, 435–450.

40.

Legendre, P. & Fortin, M.-J. (2010). Comparison of the Mantel test and alternative approaches for detecting complex multivariate relationships in the spatial analysis of genetic data. *Mol. Ecol. Resour.*, 10, 831–844.

41.

Legendre, P. & Gallagher, E.D. (2001). Ecologically meaningful transformations for ordination of species data. *Oecologia*, 129, 271–280.

42.

Legendre, P. & Legendre, L. (2012). *Numerical ecology*. 3rd English edition. Elsevier Science BV, Amsterdam.

43.

Nekola, J.C. & White, P.S. 1999). The distance decay of similarity in biogeography and ecology. *J. Biogeogr.*, 26, 867–878.

44.

Odum, E.P. (1950). Bird populations of the Highlands (North Carolina) Plateau in relation to plant succession and avian invasion. *Ecology*, 31, 587–605.

45.

Oksanen, J., Blanchet, F.G., Kindt, R., Legendre, P., Minchin, P.R., O'Hara, R.B. *et al.* (2012). *vegan: Community ecology package. R package version 2.0-3.* Available at: http://cran.r-project.org/web/packages/vegan/.

778 46.
779 Orlóci, L. (1967). An agglomerative method for classification of plant communities. *J. Ecol.*, 55,
780 193–206.
781 47.
782 Orlóci, L. (1978). *Multivariate analysis in vegetation research.* 2nd edition. Dr. W. Junk B. V.,
783 The Hague, The Netherlands.
784 48.
785 Pelissier, R., Couteron, P., Dray, S. & Sabatier, D. (2003). Consistency between ordination
786 techniques and diversity measurements: two strategies for species occurrence data.
787 *Ecology*, 84, 242–251.
788 49.
789 Peres-Neto, P.R., Legendre, P., Dray, S. & Borcard, D. (2006). Variation partitioning of species
790 data matrices: estimation and comparison of fractions. *Ecology*, 87, 2614–2625.
791 50.
792 Rao, C.R. (1982). Diversity and dissimilarity coefficients: a unified approach. *Theor. Popul.*
793 *Biol.*, 21, 24–43.
794 51.
795 Rao, C.R. (1995). A review of canonical coordinates and an alternative to correspondence
796 analysis using Hellinger distance. *Qüestiió (Quaderns d'Estadistica i Investivació*
797 *Operativa)*, 19, 23–63.
798 52.
799 Ricotta, C. & Marignani, M. (2007). Computing B-diversity with Rao's quadratic entropy: a
800 change of perspective. *Divers. Distrib.*, 13, 237–241.

801 53.

802 Stephenson, W., Williams, W.T. & Cook, S.D. (1972). Computer analyses of Petersen's original
803 data on bottom communities. *Ecol. Monogr.*, 42, 387–415.

804 54.

805 Vellend, M. (2001). Do commonly used indices of beta-diversity measure species turnover? *J.*
806 *Veg. Sci.*, 12, 545–552.

807 55.

808 Verneaux, J. (1973). Cours d'eau de Franche-Comté (Massif du Jura). Recherches écologiques
809 sur le réseau hydrographique du Doubs – Essai de biotypologie. *Annales Scientifiques de*
810 *l'Université de Franche-Comté, Biologie Animale*, 3, 1–260.

811 56.

812 Wagner, H.H. (2003). Spatial covariance in plant communities: integrating ordination, variogram
813 modeling, and variance testing. *Ecology*, 84, 1045–1057.

814 57.

815 Wagner, H.H. (2004). Direct multi-scale ordination with canonical correspondence analysis.
816 *Ecology*, 85, 342–351.

817 58.

818 Whittaker, R.H. (1952). A study of summer foliage insect communities in the Great Smoky
819 Mountains. *Ecol. Monogr.*, 22, 1–44.

820 59.

821 Whittaker, R.H. (1960). Vegetation of the Siskiyou mountains, Oregon and California. *Ecol.*
822 *Monogr.*, 30, 279–338.

823 60.

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

61.

Wilson, M.V. & Shmida, A. (1984). Measuring beta diversity with presence-absence data. *J. Ecol.*, 72, 1055–1064.

62.

Wishart, D. (1969). *CLUSTAN 1a user manual.* Computing Laboratory, University of St. Andrews, St. Andrews, Fife, Scotland.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article:

Appendix S1 Community composition data transformations.

Appendix S2 Computing the total sum of squares from a dissimilarity matrix.

Appendix S3 Details about the properties of dissimilarity coefficients.

Appendix S4 The R function beta.div() computes estimates of total beta diversity as the total variance in a community data table **Y**, as well as the derived SCBD and LCBD statistics, for 16 dissimilarity coefficients or the raw data table.

Table 1 Dissimilarity coefficients compared in this paper.

Dissimilarity	Abundance-based	Incidence-based	References	Coefficient no. in L&L [1]
Euclidean distance	$\sqrt{\sum_{j=1}^{p}\left[y_{1j}-y_{2j}\right]^2}$	$\sqrt{p\left(\frac{b+c}{a+b+c+d}\right)}=\sqrt{b+c}$		D_1
Manhattan distance	$\sum_{j=1}^{p}\lvert y_{1j}-y_{2j}\rvert$	$p\left(\frac{b+c}{a+b+c+d}\right)=b+c$		D_7
Modified mean character difference	$\frac{1}{pp}\sum_{j=1}^{p}\lvert y_{1j}-y_{2j}\rvert$	$\frac{b+c}{a+b+c}$	Legendre & Legendre (2012)	D_{19}
Species profile distance	$\sqrt{\sum_{j=1}^{p}\left[\frac{y_{1j}}{y_{1+}}-\frac{y_{2j}}{y_{2+}}\right]^2}$	$\sqrt{\frac{b+c}{(a+b)(a+c)}}$	Legendre & Gallagher (2001)	D_{18}
Hellinger distance	$\sqrt{\sum_{j=1}^{p}\left[\sqrt{\frac{y_{1j}}{y_{1+}}}-\sqrt{\frac{y_{2j}}{y_{2+}}}\right]^2}$	$\sqrt{2\left(1-\frac{a}{\sqrt{(a+b)(a+c)}}\right)}$	Rao (1995)	D_{17}
Chord distance	$\sqrt{\sum_{j=1}^{p}\left[\frac{y_{1j}}{\sqrt{\sum_{k=1}^{p}y_{1k}^2}}-\frac{y_{2j}}{\sqrt{\sum_{k=1}^{p}y_{2k}^2}}\right]^2}$	$\sqrt{2\left(1-\frac{a}{\sqrt{(a+b)(a+c)}}\right)}$	Orlóci (1967)	D_3
Chi-square distance	$\sqrt{y_{++}\sum_{j=1}^{p}\frac{1}{y_{+j}}\left[\frac{y_{1j}}{y_{1+}}-\frac{y_{2j}}{y_{2+}}\right]^2}$	NA [2]	Lebart & Fénelon (1971)	D_{16}
Coefficient of divergence	$\sqrt{\frac{1}{pp}\sum_{j=1}^{p}\left(\frac{y_{1j}-y_{2j}}{y_{1j}+y_{2j}}\right)^2}$	$\sqrt{\frac{b+c}{a+b+c}}$	Clark (1952)	D_{11}
Canberra metric [3]	$\frac{1}{pp}\sum_{j=1}^{p}\frac{\lvert y_{1j}-y_{2j}\rvert}{\left(y_{1j}+y_{2j}\right)}$	$\frac{b+c}{a+b+c}$	Lance & Willams (1967), Stephenson *et al.* (1972) for 1/pp	D_{10}

Whittaker's index of association	$\frac{1}{2}\sum_{j=1}^{p}\left\lvert\frac{y_{1j}}{y_{1+}}-\frac{y_{2j}}{y_{2+}}\right\rvert$	$\frac{1}{2}\left(\frac{b}{a+b}+\frac{c}{a+c}+\left\lvert\frac{a}{a+b}-\frac{a}{a+c}\right\rvert\right)$	Whittaker (1952)	D_9
Percentage difference (*alias* Bray-Curtis dissimilarity [4])	$\frac{\sum_{j=1}^{p}\lvert y_{1j}-y_{2j}\rvert}{y_{1+}+y_{2+}}$	$\frac{b+c}{2a+b+c}$	Odum (1950)	D_{14}
Wishart coefficient = (1 – similarity ratio)	$1-\left[\frac{\sum_{j=1}^{p}y_{1j}y_{2j}}{\sum_{j=1}^{p}y_{1j}^2+\sum_{j=1}^{p}y_{2j}^2-\sum_{j=1}^{p}y_{1j}y_{2j}}\right]$	$\frac{b+c}{a+b+c}$	Wishart (1969), Janssen (1975)	
D = (1–Kulczynski coefficient)	$1-\frac{1}{2}\left[\frac{\sum_{j=1}^{p}\min(y_{1j},y_{2j})}{y_{1+}}+\frac{\sum_{j=1}^{p}\min(y_{1j},y_{2j})}{y_{2+}}\right]$	$1-\frac{1}{2}\left(\frac{a}{a+b}+\frac{a}{a+c}\right)$	Kulczynski (1928)	$1 - S_{18}$
Abundance-based Jaccard [5]	$\left(1-\frac{UV}{U+V-UV}\right)$	$\frac{b+c}{a+b+c}$	Chao *et al.* (2006)	
Abundance-based Sørensen [5]	$\left(1-\frac{2UV}{U+V}\right)$	$\frac{b+c}{2a+b+c}$	Chao *et al.* (2006)	
Abundance-based Ochiai [5]	$\left(1-\sqrt{UV}\right)$	$\left(1-\frac{a}{\sqrt{(a+b)(a+c)}}\right)$	Chao *et al.* (2006)	

[1]L&L = Legendre & Legendre (2012).

[2]NA: No binary form with parameters *a*, *b* and *c* for this coefficient, although it can be computed for presence-absence data.

[3]Division by *pp* (number of species excluding double zeros) introduced by Stephenson *et al.* (1972) and adopted by Oksanen *et al.* (2012).

[4]Coefficient first described by Steinhaus in the 1940's, then by Odum (1950) as the *percentage difference*. The Bray & Curtis (1957) paper described a new ordination method; the index described and used by these authors was Whittaker's dissimilarity, not the percentage difference which is more general. It is incorrect to attribute this coefficient to these authors.

[5]*U* and *V* notation: see Chao *et al.* (2006).

Table 2 Properties P4 to P14 of the coefficients in Table 1. P1 to P3 (not shown) are fulfilled by all coefficients. Property descriptions are found in Appendix S3. 1 indicates that a coefficient has the property, 0 that it does not. For P13, code 2 indicates that both **D** and $\mathbf{D}^{(0.5)}$ are Euclidean, 1 that only $\mathbf{D}^{(0.5)} = [D_{hi}^{0.5}]$ is Euclidean, and 0 that neither **D** nor $\mathbf{D}^{(0.5)}$ are Euclidean. NA: there is no binary form for the chi-square distance, hence P10 could not be assessed. Last column: maximum possible dissimilarity value (D_{max}) when it exists. P1 to P9 are essential properties for beta assessment; P10 to P14 describe additional properties, useful for special applications.

Dissimilarity	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	D_{max}
Euclidean distance	0	0	1	0	0	0	0	0	0	2	1	—
Manhattan distance	0	0	1	0	0	0	0	0	0	1	0	—
Modified mean character difference	1	0	1	1	0	0	1	0	0	0	0	—
Species profile distance	1	0	0	0	1	1	0	1	0	2	1	$\sqrt{2}$
Hellinger distance	1	1	1	1	1	1	1	1	0	2	1	$\sqrt{2}$
Chord distance	1	1	1	1	1	1	1	1	0	2	1	$\sqrt{2}$
Chi-square distance	1	0	1	1	1	1	NA	0	0	2	1	$\sqrt{2y_{++}}$
Coefficient of divergence	1	1	1	1	1	1	1	0	0	2	0	1

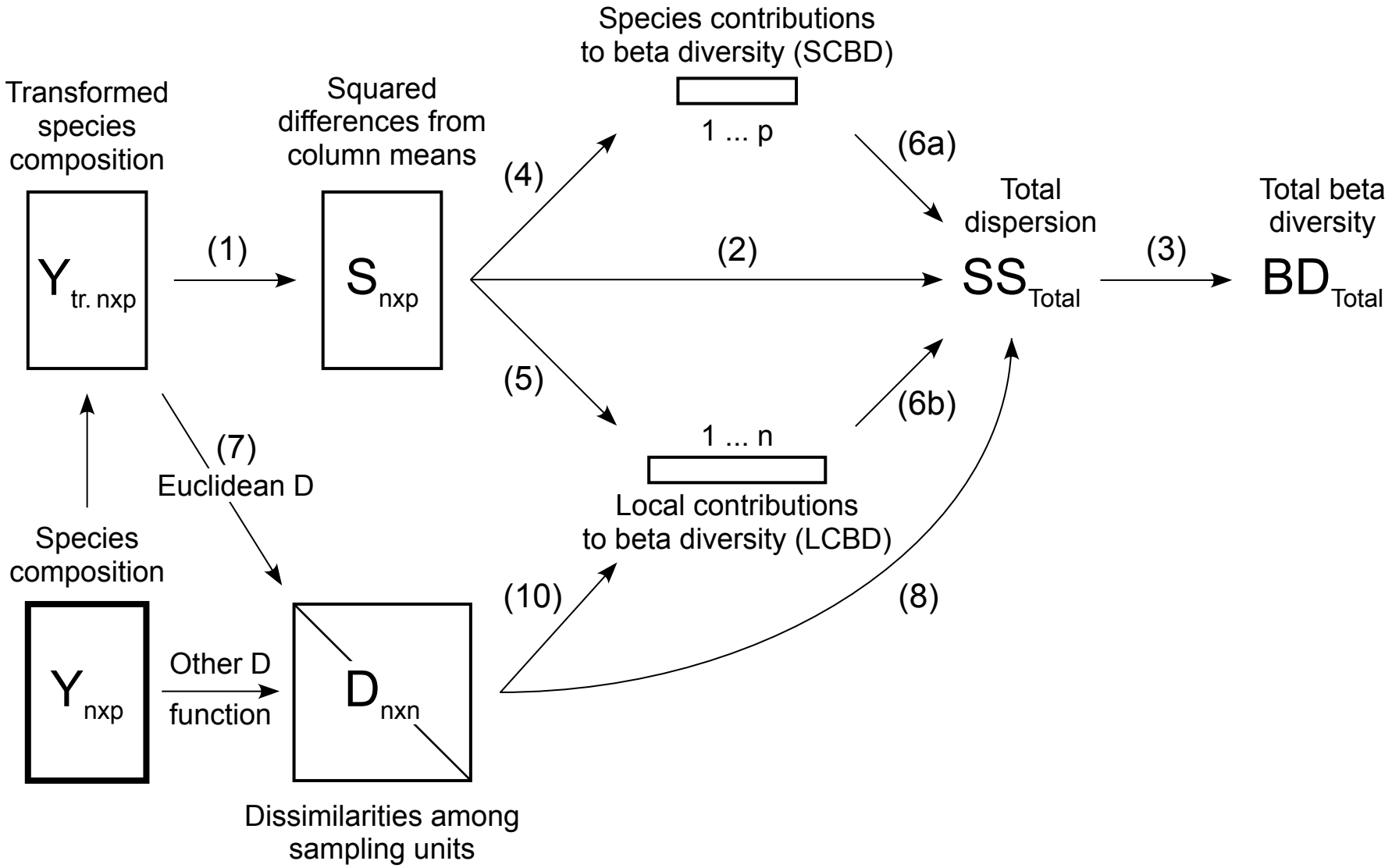
Canberra metric	1	1	1	1	1	1	1	0	0	1	0	1
Whittaker's index of association	1	1	1	1	1	1	1	1	0	1	0	1
Percentage difference (*alias* Bray-Curtis)	1	1	1	1	1	1	1	0	0	1	0	1
Wishart coefficient = (1 – similarity ratio)	1	1	1	1	1	1	1	0	0	1	0	1
D = (1–Kulczynski coefficient)	1	1	1	1	1	1	1	0	0	0	0	1
Abundance-based Jaccard	1	1	1	1	1	1	1	1	1	0	0	1
Abundance-based Sørensen	1	1	1	1	1	1	1	1	1	0	0	1
Abundance-based Ochiai	1	1	1	1	1	1	1	1	1	0	0	1

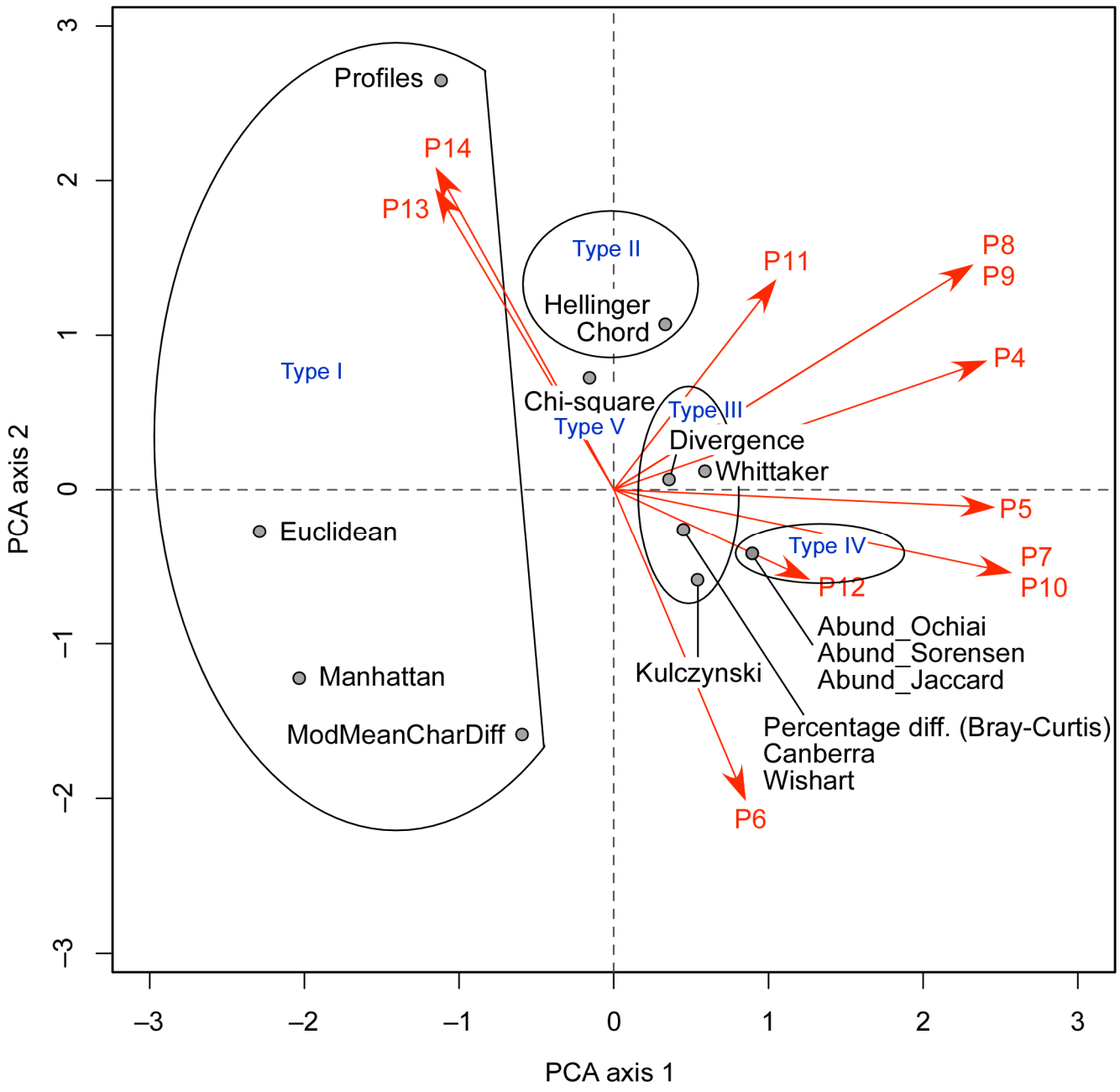
Figure 1 Schematic diagram representing the different ways of computing beta diversity as the total variance in the species composition data table **Y**, as well as the contributions of individual species and sampling units. Numbers in parentheses refer to equations in the text.

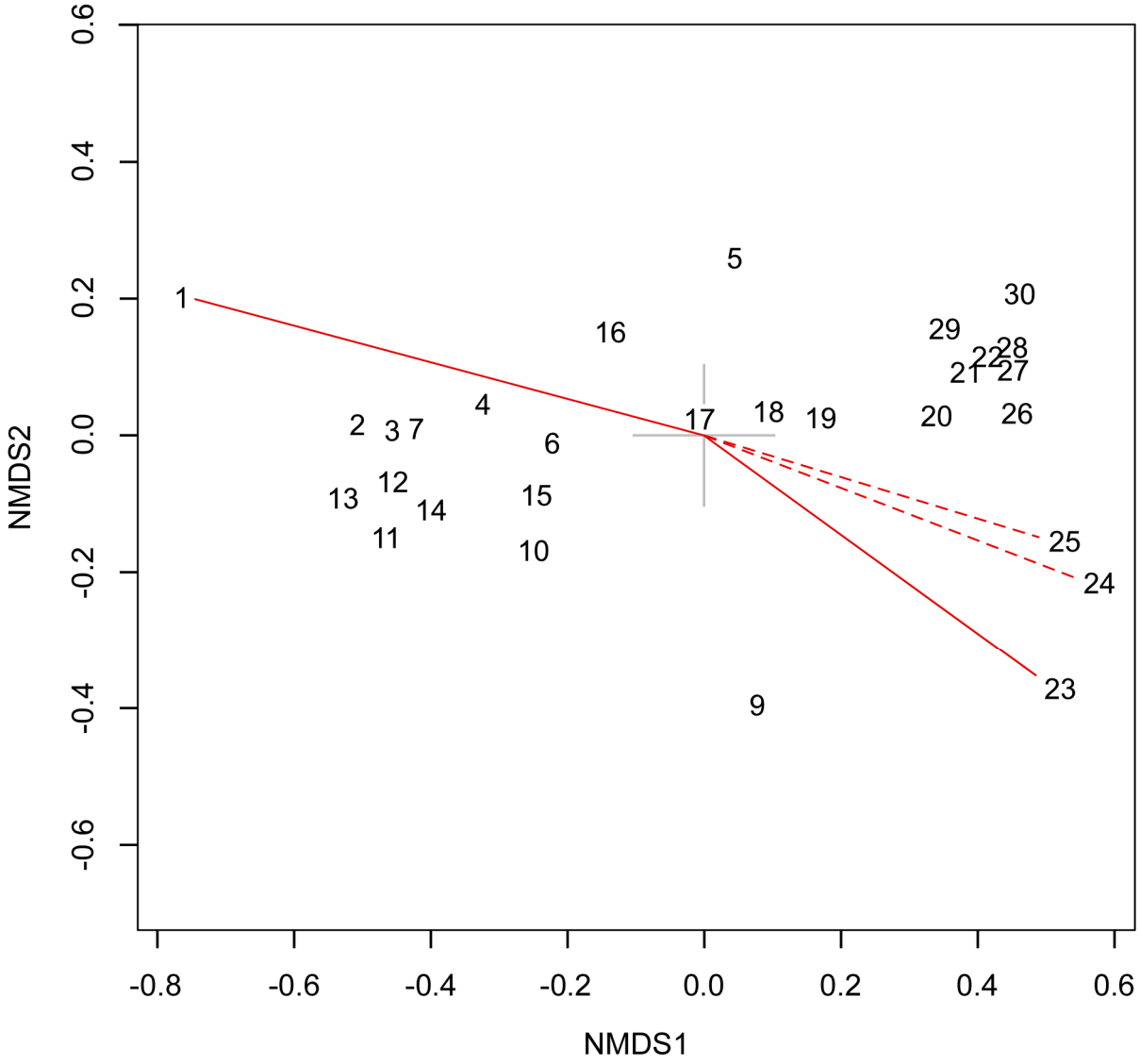
Figure 2 Principal component biplot relating properties P4 to P14 (red arrows) to the dissimilarity coefficients (gray points; see Table 1 for the full coefficient names). The five types of coefficients (blue labels), shown in the figure, are described in the text. PCA axis 1 accounts for 46% of the multivariate variation and axis 2 for 23%.

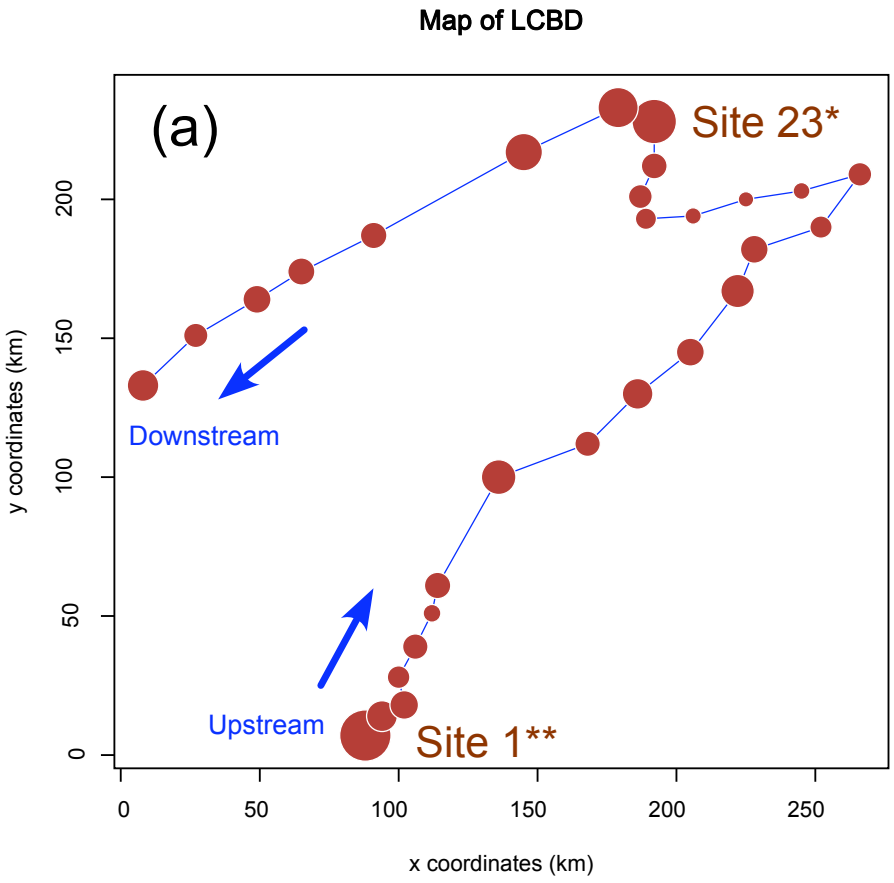
Figure 3 Ordination diagram of Doubs River fish data sites (nonmetric multidimensional scaling, nMDS; chord distance). SS_i indices are the squares of the distances of the sites to the multivariate centroid. The significant indices ($p < 0.05$) are represented by red lines joining the points to the centroid (full lines: $p < 0.05$ after Holm correction for 29 simultaneous tests).

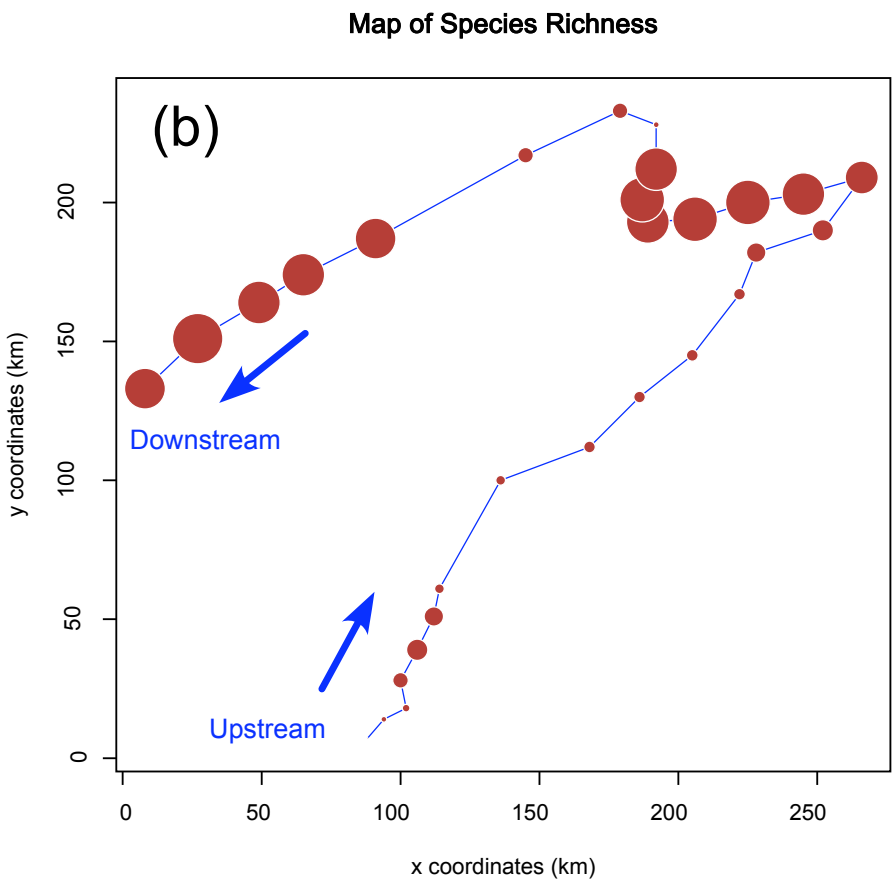
Figure 4 Maps of Doubs River (blue line) showing (a) the local contributions to beta diversity (LCBD) of the fish assemblage data and (b) the species richness at the 29 study sites. Size of the circles is proportional to the LCBD or richness values. Two sites have significant LCDB (or SS_i) indices at the 0.05 significance level after Holm correction for multiple testing: site 1 ($p = 0.003$) and site 23 ($p = 0.042$). The arrows indicate flow direction.


Species contributions
to beta diversity (SCBD)
1 ... p
Transformed
species
composition
Squared
differences from
column means
(4)
(6a)
Total
dispersion
Total beta
diversity
Y tr. nxp
(1)
S nxp
(2)
SS Total
(3)
BD Total
(5)
(6b)
1 ... n
(7)
Euclidean D
Local contributions
to beta diversity (LCBD)
Species
composition
(10)
(8)
Y nxp
Other D
function
D nxn
Dissimilarities among
sampling units


Profiles
P14
P13
Type II
Hellinger
Chord
P11
P8
P9
P4
Type I
Chi-square
Type V
Type III
Divergence
Whittaker
P5
Euclidean
Type IV
P7
P10
P12
Abund_Ochiai
Abund_Sorensen
Abund_Jaccard
Manhattan
Kulczynski
Percentage diff. (Bray-Curtis)
Canberra
Wishart
ModMeanCharDiff
P6
PCA axis 2
PCA axis 1
3
2
1
0
−1
−2
−3
−3
−2
−1
0
1
2
3


NMDS2
NMDS1
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.8
-0.6
-0.4
-0.2
0.0
0.2
0.4
0.6
1
2
3
7
4
6
5
16
17
18
19
20
29
30
28
22
21
27
26
13
12
14
11
15
10
25
24
23
9


Map of LCBD
(a)
Site 23*
Downstream
Upstream
Site 1**
y coordinates (km)
x coordinates (km)
0
50
100
150
200
250


Map of Species Richness
(b)
Downstream
Upstream
y coordinates (km)
x coordinates (km)
0
50
100
150
200
250

Appendix S1

COMMUNITY COMPOSITION DATA TRANSFORMATIONS

The following data transformations (Legendre & Gallagher 2001), applied to species frequency data (or frequency-like data such as biomass) before computing the Euclidean or Manhattan distance, produce distance coefficients that are included in our comparative study:

• Species profile transformation: $y'_{ij} = y_{ij} / y_{i+}$;

• Hellinger transformation: $y'_{ij} = \sqrt{y_{ij} / y_{i+}}$;

• Chord transformation: $y'_{ij} = y_{ij} \Big/ \sqrt{\sum_{j=1}^{p} y_{ij}^2}$;

• Chi-square transformation:

$$y'_{ij} = \sqrt{y_{++}} \frac{y_{ij}}{y_{i+}\sqrt{y_{+j}}} \quad \text{where } y_{i+} = \sum_{j=1}^{p} y_{ij}, \quad y_{+j} = \sum_{i=1}^{n} y_{ij}, \text{ and } y_{++} = \sum_{i=1}^{n}\sum_{j=1}^{p} y_{ij}.$$

After computation of the Euclidean distance, the corresponding dissimilarities are the species profile, Hellinger, chord, and chi-square distances. The Hellinger and chord distances are appropriate for beta diversity studies, but the distance between species profiles and the chi-square distance are not; see *Comparative study* in the main paper.

Before calculation of the Euclidean or Manhattan distance, another approach is to transform community composition data using simple transformations. Examples are the usual square-root and $y'_{ij} = \log(y_{ij} + c)$ transformations (constant c is usually 1 when transforming species frequency data, but it could take other values for biomass data for example), or the special log transformation of Anderson *et al.* (2006), which makes allowance for species frequencies of

zeros. Log transformations are appropriate for species data with log-normal distribution; log-transformed data can then be used as input into the percentage difference and Kulczynski dissimilarities. Other transformations that are appropriate for community composition data were described by Faith *et al.* (1987), among other authors.

The Euclidean distance computed on data transformed using the square-root, $\log(y_{ij} + c)$, or Anderson's log transformations still lacks properties P4, P5, P7, P8 and P9 that are essential for beta diversity assessment (Appendix S3). These transformations do not solve the problems of the Euclidean distance computed on raw abundance data (Table 2).

Community composition data transformed following any of the transformations described in this section can be used in linear models such as simple (PCA) and canonical (RDA) ordination, *K*-means partitioning, and multivariate regression tree analysis (MRT); these methods implicitly preserve the Euclidean distance among sites.

Computing the Manhattan distance on data transformed into species profiles produces Whittaker's index of association multiplied by 2, which is an appropriate coefficient for beta diversity studies (Whittaker 1952). The Manhattan distance is, however, not the distance implicit in linear models, so that Whittaker's index of association does not lend itself to linear modelling nor to the calculation of *Species Contributions to Beta Diversity* (SCDB indices) described in the main paper, eqn. 4b.

REFERENCES

Anderson, M.J., Ellingsen, K.E. & McArdle, B.H. (2006). Multivariate dispersion as a measure of beta diversity. *Ecol. Lett.*, 9, 683–693.

Faith, D.P., Minchin, P.R. & Belbin, L. (1987). Compositional dissimilarity as a robust measure of ecological distance. *Vegetatio*, 69, 57–68.

Legendre, P. & Gallagher, E.D. (2001). Ecologically meaningful transformations for ordination of species data. *Oecologia*, 129, 271–280.

Whittaker, R.H. (1952). A study of summer foliage insect communities in the Great Smoky Mountains. *Ecol. Monogr.*, 22, 1–44.

Appendix S2

COMPUTING THE TOTAL SUM OF SQUARES FROM A DISSIMILARITY MATRIX

Equation 8 in the main paper shows how to compute the total sum of squares, SS_{Total}, for a matrix of Euclidean distances. The equivalence of eq. 2 (computed on raw data) and eq. 8 (computed on distances) was demonstrated in Appendix 1 of Legendre & Fortin (2010). The profile, chord, Hellinger, and chi-square distances are obtained by first transforming the raw abundance data as described in Appendix S1, then computing the Euclidean distance formula on the transformed data. As a consequence, for these distances computed using the Euclidean distance formula, it is clear that SS_{Total} can be computed either from the transformed data through eq. 2 or from the distances through eq. 8.

For the other distances that have the Euclidean property as either **D** or $\mathbf{D}^{(0.5)} = [D_{hi}^{0.5}]$ (Table 2, column P13, codes 1 or 2), eq. 8 also applies. That point is demonstrated as follows: any dissimilarity matrix that has the Euclidean property can be decomposed into principal coordinates by principal coordinate analysis (PCoA, Gower 1966), obtaining a fully Euclidean representation of the data (Gower & Legendre 1986). Calculation of dissimilarity matrix **D** followed by PCoA of **D** or $\mathbf{D}^{(0.5)}$ acts as a data transformation. The total sum of squares of the matrix of principal coordinates, computed through eq. 2, is equal to the total sum of squares computed from dissimilarity matrix **D** or $\mathbf{D}^{(0.5)}$ through eq. 8.

Finally, for dissimilarities that do not lead to a fully Euclidean representation (i.e. those that do not have the Euclidean property; Table 2, column P13, code 0), the same equivalence still exists although these matrices produce negative eigenvalues in PCoA. The demonstration involves two steps. On the one hand, the trace of the Gower-centred matrix on which eigen-

decomposition is computed, which is equal to the sum of all eigenvalues (positive and negative), is equal to SS_{Total} computed from the dissimilarity matrix through eq. 8. On the other hand, McArdle & Anderson (2001) and Anderson (2006) have shown how to compute SS_{Total} of the principal coordinate representation using the real and complex principal coordinates, and this is equal again to the trace of the Gower-centred matrix on which eigen-decomposition is computed in PCoA. Their method has three steps: (1) square all values in the matrix of eigenvectors (which were produced by eigen-decomposition with a norm of 1), (2) multiply each squared eigenvector by its eigenvalue, and (3) sum all the resulting values. The calculation is demonstrated in the R function pcoa.short() in R, below.

Illustration: calculation of SS_{Total} for a non-Euclidean dissimilarity matrix

For the example, we will use the first 10 rows of the mite data available in package **vegan** and compute the percentage difference dissimilarity. The calculations are done in the R language.

```
require(vegan)
data(mite)

### Compute the percentage difference dissimilarity for the mite data (first 10 rows)
mite.D <- vegdist(mite[1:10,], "bray")
# Is the dissimilarity matrix Euclidean?
require(ade4)
is.euclid(mite.D)                                   # Available in R package ade4
# [1] FALSE

### Compute SSTotal from the dissimilarities

SS.D <- function(D, n) sum(D^2) / (n)               # Equation 8
res.SS.D <- SS.D(mite.D, 10)
res.SS.D
# [1] 0.9626073                                     # Result: SSTotal

### A short function for principal coordinate analysis (PCoA)

#############################################################################
pcoa.short <- function(D, include.zero=FALSE, only.values=FALSE)
#
# Compute PCoA for a Euclidean or non-Euclidean dissimilarity matrix.
```

```
# The eigenvectors are not scaled to sqrt(eigenvalues),
# hence they are not principal coordinates in the PCoA sense.
#
# When 'n' is very large, users may choose not to compute the eigenvectors.
# This is obtained by selecting the option only.values=TRUE.
# The statistic VarTotal=SS will not be computed and printed in that case.
#
# License: GPL-2
# Author:: Pierre Legendre
{
	D <- as.matrix(D)
	n <- nrow(D)
	epsilon <- sqrt(.Machine$double.eps)
#
# Gower centring, matrix formula
	One <- matrix(1,n,n)
	mat <- diag(n) - One/n
	G <- -0.5 * mat %*% (D^2) %*% mat
	trace <- sum(diag(G))
	SSi <- diag(G)
#
# Eigenvalue decomposition
	eig <- eigen(G, symmetric=TRUE, only.values=only.values)
# Exclude the null eigenvalue/s if include.zero is FALSE
	select <- 1:n
	exclude <- which(abs(eig$values) < epsilon)
	cat("Note - Eigenvalue/s", exclude, "is/are null\n")
	if(!include.zero) {
		cat("Note - Eigenvalue/s and eigenvector/s", exclude, "was/were excluded\n")
		select <- select[-exclude]
		}
	values <- eig$values[select]
#
if(!only.values) {				# Compute SS from the eigenvectors
	vectors <- eig$vectors[,select]
	vectors.sq <- vectors^2 %*% diag(values)
	SS <- sum(vectors.sq)
	} else {
	vectors <- NA
	SS <- NA
	}
#
list(values=values, vectors=vectors, trace=trace, SS.total=SS, SSi=SSi, site.names=rownames(D),
select=select)
}
##############################################################################
```

```
res <- pcoa.short(mite.D)
# Note - Eigenvalue/s 9 is/are null
# Note - Eigenvalue/s and eigenvector/s 9 was/were excluded

### Compute SS_Total as the trace of the Gower-centred matrix

res$trace
# [1] 0.9626073                                   # Result: SS_Total

### Compute SS_Total as the sum of the PCoA eigenvalues

sum(res$values)
# [1] 0.9626073                                   # Result: SS_Total

### Compute SS_Total as the sum of squares of the principal coordinates scaled to lengths equal to
the square roots of the eigenvalues, including the one with a negative eigenvalue. The result is the
same with options include.zero=FALSE or include.zero=TRUE.

res$SS.total
# [1] 0.9626073                                   # Result: SS_Total
```

REFERENCES

Anderson, M.J. (2006). Distance-based tests for homogeneity of multivariate dispersions. *Biometrics*, 62, 245–253.

Gower, J.C. & Legendre, P. (1986). Metric and Euclidean properties of dissimilarity coefficients. *J. Classif.*, 3, 5–48.

Legendre, P. & Fortin, M.-J. (2010). Comparison of the Mantel test and alternative approaches for detecting complex multivariate relationships in the spatial analysis of genetic data. *Mol. Ecol. Resour.*, 10, 831–844.

McArdle, B.H. & Anderson, M.J. (2001). Fitting multivariate models to community data: a comment on distance-based redundancy analysis. *Ecology*, 82, 290–297.

Appendix S3

DETAILS ABOUT THE PROPERTIES OF DISSIMILARITY COEFFICIENTS

We describe four groups of properties and indicate the reason why we consider them relevant. The first two groups (i.e. from P1 to P9) contain the minimum requirements for assessing beta diversity. The remaining two groups (i.e. P10 to P14) are not necessarily required in all beta diversity assessments. Practitioners should determine whether the context of their analyses requires these latter properties or not. Similarity coefficients should be transformed into dissimilarities before assessing the following properties.

Property class 1: Basic necessary properties. — Properties P1 to P6 must be fulfilled by all resemblance coefficients used for beta diversity assessment. P1 and P2 are actually mathematical axioms that define a dissimilarity function. Thus, they are fulfilled by all coefficients considered in this paper. P3 (monotonicity) is a necessary condition for any coefficient used to study species assemblages. All coefficients investigated in the present study are monotonic. Properties P1 to P3 are therefore not shown in Table 1 of the paper.

P1 – Minimum of zero and positiveness. A dissimilarity value should never be negative and it should be zero when comparing a site to itself. When comparing two different sites, it can be zero or greater than zero, depending on the species abundance values and how the dissimilarity is defined. For example, with some coefficients, D is zero when comparing two site vectors whose abundance values are proportional to each other; that is the case with the profile, chi-square, chord, and Hellinger distances. Dissimilarities that violate this property by taking negative values are called nonmetric, by opposition to the metric and semimetric coefficients (see Legendre & Legendre 2012).

P2 – Symmetry. Consider two community abundance vectors, $\mathbf{x}_1$ and $\mathbf{x}_2$, whose dissimilarity is to be assessed. In symmetric indices, $D(\mathbf{x}_1,\mathbf{x}_2) = D(\mathbf{x}_2,\mathbf{x}_1)$. In the incidence-based counterparts of these coefficients (Table 1 in the main paper), the values *b* and *c* play exchangeable (symmetric) roles. When studying beta diversity, there is no reason to make a distinction between the two sampling units that are compared using a coefficient. Therefore, dissimilarity coefficients must be symmetric. The property of being *double-zero symmetrical*, referred to in P4, is different.

P3 – Monotonicity to changes in abundance. Increasing the difference in abundances of one of several species between two sites increases their dissimilarity. Property P3 was verified using ordered comparison case series (OCCAS), corresponding to linear changes in the abundances of two species along different types of simulated environmental gradients. The OCCAS method was proposed by Hajdu (1981) and used by Gower & Legendre (1986) to assess monotonicity in dissimilarity coefficients.

P4 – Double-zero asymmetry. Coefficients that have this property do not change when double-zeros are added to the data, but the dissimilarity decreases when double-X (where X > 0) values are added. The reasoning implies two conditions, derived from ecological niche theory.

1. Ecological statement: double-zeros in species abundance are not interpretable. — In his seminal paper, Whittaker (1972) published a figure (his Fig. 4, p. 228) showing simulated species represented by bell-shaped curves with different widths (species tolerances), succeeding one another along three ecological gradients. (1.1) For species *j* observed at a pair of sites, the presence of that species (in any abundance) at both sites indicates that the two sites are similar to some extent, i.e. they are close in positions along the gradient. Because the two sites are within the tolerance zone of species *j*, that species can be found at these two sites. (1.2) Ruling out sampling error, the presence at one site and absence at the other unambiguously indicates that the

sites occupy different positions along the gradient. (1.3) Ruling out sampling error again, double absence of species *j* is not interpretable, because it can result from the two sites being either at close positions along the gradient but outside the tolerance zone of species *j* (e.g. pH too high at both sites for that species), or at positions far away along the gradient, both sites being outside the tolerance zone of species *j* (e.g. pH too low at one site and too high at the other).

2. Lemma. — The presence of species *j* at two sites (point 1.1 above) is an indication of resemblance of these sites whatever the abundances observed. It follows that presence of species *j* with the same abundance X at two sites (a difference (X – X) = 0 for that species; point 1.1 above) has a different meaning from the double-absence (which also corresponds to a difference (0 – 0) = 0; point 1.3 above).

3. This reasoning leads to the following conclusions. (3.1) Double-zeros (0, 0) have a different meaning than double-presences (X, X), whatever the abundances. (3.2) Coefficients that produce the same effect (i.e. no change in dissimilarity) for double zeros as they do for double presences with identical abundances (i.e. (X,X), which we call double-X), where X > 0) are called *double-zero symmetrical* because they treat double zeros like any other pair of identical values. These coefficients are not admissible for the study of ecological differentiation of communities, i.e. for beta diversity studies. The Euclidean and Manhattan distances belong to that type: double-zeros and double-X produce no change in distance, whereas any other pair of non-identical abundances does produce a change in the distance. (3.3) Coefficients useful for beta diversity studies must be *double-zero asymmetrical* (term used in Legendre and Legendre 2012), meaning that their value does not change with the addition of double zeros, but it decreases when species with double-X abundances that are not double-zeros are added to the comparison of two sites.

The difference between double-zero and double-X is clearly recognized in binary coefficients. In double-zero symmetrical similarity coefficients such as the simple matching index, the numbers of double-presences and of double-absences (usually respectively represented by *a* and *d*) are used as indications of similarity. In contrast, double-zero asymmetrical similarity coefficients such as the Jaccard and Sørensen indices exclude double-absences (*d*) from both their numerator and denominator and use only the double-presences (*a*) in their formula in addition to *b* and *c*. Some binary dissimilarity coefficients can be expressed as function of *b* and *c* only (see, for example, the presence-absence formulas for the Euclidean and Manhattan distances in Table 1 of the main paper). These are also double-zero symmetrical, because neither double-presences (*a*) nor double-absences (*d*) can change the value of the coefficient. A double-zero asymmetrical binary dissimilarity coefficient must include double-presences (*a*) and exclude double-absences (*d*) from its formula.

P5 – Sites without species in common have the largest dissimilarity. Coefficients that violate this property are not suitable for beta diversity studies. For the coefficients that have a fixed upper bound (last column of Table 2), the largest dissimilarity is the upper bound. The following data sets were used to test the 16 coefficients analysed in this paper. In each example, the rows are sites S1-S4, the columns represent species.

Example data 1 –

	1	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***	***11***	***12***	***13***	***14***	***15***	***16***	***17***	***18***	***19***	***20***
S1	3	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
S2	7	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
S3	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0

S4	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1

Example data 2 –

	1	***2***	***3***	***4***
S1	1	4	0	0
S2	4	1	0	0
S3	0	1	1	0
S4	1	0	0	1

Example data 3 –

	1	***2***	***3***	***4***
S1	1	2	0	0
S2	2	1	0	0
S3	0	1	40	0
S4	1	0	0	40

Some dissimilarity functions present the following paradox with one of more of these data sets: for sites 3 and 4 that have no species in common, these coefficients produce dissimilarities that are smaller than for sites that have some (e.g. sites 1 and 3) or all (sites 1 and 2) their species in common. These examples are extensions of the Orlóci (1978, p. 46) paradox data.

P6 – Dissimilarity does not decrease in series of nested species assemblages. For pairs of sites having any number of unique species, i.e. species that are not found at the other site, the dissimilarity should be the same or increase with the number of unique species. In particular, the dissimilarity should not decrease when the number of unique species in one or both sites increases. To assess this property, we carried out simulations where we added unique species to one of the sites (which corresponds to the data structure described as *nestedness* of species assemblages by Wright & Reeves 1992 and Baselga 2010) or to both sites (data structure described as monotonic by Jost *et al.* 2011). Violation of P6 leads to the paradox that the total sum of squares for a pair of sites, which is $D^2/2$ (eq. 8 in the main paper), tends to 0 as the number of unique species increases.

Property class 2: Comparability between data sets. — The following three properties are needed to appropriately compare beta diversity values calculated for different data tables, even if the sampling unit are the same size (e.g. quadrat size for vegetation) and the sampling effort is the same. Therefore, we consider they should also be required in beta diversity studies.

P7 – Species replication invariance. A community composition table with the columns in two or several copies should produce the same dissimilarities among sites as the original data table. Procedure: select a community composition data table, compute a dissimilarity index and obtain a dissimilarity matrix. Then, duplicate the data table, combining the two copies side by side, using for example function cbind() in R. Compute the same dissimilarity function and obtain a new dissimilarity matrix. Repeat the duplication step to include three copies of the data and compute the dissimilarity matrix again. The three dissimilarity matrices should be identical if the dissimilarity function has the property of *replication invariance*. For the abundance-based coefficients, the property is computed using the population formula, not the sample formula. This

property was first described by Jost *et al.* (2011).

A special case of this property (not included as a separate property in the present study), for presence-absence data, is called *homogeneity*, for example by Janson and Vegelius (1981), Koleff *et al.* (2003) and Chao *et al.* (2006). The homogeneity property allows the comparison of beta values computed from data tables containing different total species richness. This property is verified on the binary form of the coefficients, by multiplying a, b, c and d by a constant factor and checking whether the resulting index value is changed.

P8 – Invariance to the measurement units. This property concerns abundance-based formulas only. It allows the comparison of beta values between data tables (e.g. regions) with different productivities (abundance or biomass), or where biomass has been measured using different units (e.g. in g and mg). To see whether a given quantitative coefficient is invariant to changes of measurement scale, we multiplied the abundance values by a constant factor and checked whether the resulting index was altered.

P9 – Existence of a fixed upper bound. The existence of an upper bound for a coefficient facilitates the interpretation and comparison of beta values because an upper bound in the dissimilarity index leads to an upper bound in the beta diversity value. The maximum beta value for a region is obtained when all site pairs have the maximum dissimilarity D_{max} permitted by the chosen coefficient. One can apply eq. 8 to that situation to compute the maximum sum of squares:

$$SS_{max} = \frac{1}{n}\left(\frac{n(n-1)}{2} D_{max}^2\right) = \frac{n-1}{2} D_{max}^2$$

then eq. 3 to obtain the maximum beta diversity value:

139 $$BD_{max} = SS_{max}/(n-1) = \left(\frac{n-1}{2} D^2_{max}\right)\frac{1}{n-1} = \frac{1}{2} D^2_{max}$$

140 The upper bound varies among dissimilarity coefficients (Table 2 of the pain paper, right-hand
141 column). For coefficients with $D_{max} = \sqrt{2}$, $BD_{max} = 1$; for those with $D_{max} = 1$, $BD_{max} = 0.5$ (see
142 section "Maximum value of BD" in the pain paper). For the chi-square distance, $D_{max} = \sqrt{2y_{++}}$
143 and $BD_{max} = y_{++}$ which is the sum of the species abundances in **Y**. Although y_{++} varies from data
144 table to data table, the chi-square distance is considered as belonging to the group of the
145 coefficients that have a fixed upper bound because its sister index, the chi-square metric (not
146 otherwise discussed in this paper), has an upper bound of $\sqrt{2}$ (see e.g. Legendre and Legendre
147 2012). The chi-square distance is the chi-square metric multiplied by $\sqrt{y_{++}}$, hence it has an upper
148 bound of $\sqrt{2y_{++}}$. The chi-square distance is the one computed in software packages; this is why
149 its properties are described here. The chi-square metric and distance have the same properties
150 besides their different maximum values. Hence, for coefficients that have a fixed maximum (see
151 section "The dissimilarity measures" in the main paper), we can compute a relative value of beta
152 diversity, BD_{rel}, as follows:

153 $$BD_{rel} = BD_{Total}/BD_{max}$$

154 which is a value between 0 and 1. BD_{rel} is useful to compare beta values computed using different
155 coefficients.

156 *Property class 3: Sampling issues*. — This group of properties is mostly related to sampling
157 issues. The fulfilment of properties P10 and P11 facilitates (but does not ensure) the
158 comparability of beta values obtained from sampling units having different sizes or sampled
159 using different efforts. Indeed, both the number of species and the total abundance may be

strongly affected by changes in the size of sampling units or in sampling effort. On the other hand, if the size of sampling units and sampling effort are sufficiently homogeneous, ecologists may be interested in allowing differences in the numbers of species, and perhaps also in the total abundances between sites, to influence the dissimilarity and beta diversity assessments.

The last property deals with correction for undersampling (P12) of the community composition. This property is also related to sampling effort. It is related to sampling unit size as well because small sampling units can lead to undersampling the richness of the targeted community.

P10 – Invariance to the number of species in each sampling unit. This property analyses whether a double-zero asymmetrical binary coefficient changes its value depending on the number of species in each of two sampling units $\mathbf{x}_1$ and $\mathbf{x}_2$ that are compared. Does the dissimilarity value change if the two communities are species rich, compared to when the two communities are species poor or when one is rich and the other poor?

This property was verified algebraically on the binary form of the coefficients; it could not be checked for the chi-square metric which does not have a binary form. We start with the usual a, b, c notation for binary indices:

a = number of species shared between $\mathbf{x}_1$ and $\mathbf{x}_2$

b = number of unique species in $\mathbf{x}_1$ that do not appear in $\mathbf{x}_2$

c = number of unique species in $\mathbf{x}_2$ that do not appear in $\mathbf{x}_1$

We then define the number of species in $\mathbf{x}_1$ and $\mathbf{x}_2$ as $n_1 = a + b$ and $n_2 = a + c$, and the proportion of shared species with respect to each site as $p_1 = a / n_1 = a / (a + b)$ and $p_2 = a / n_2 = a / (a + c)$. After defining n_1, n_2, p_1 and p_2, one can reformulate the binary dissimilarity measures found in column *Incidence-based* of Table 1 of the main paper in terms of these four quantities, instead of using the notation a, b, c. The idea of the proof is to see whether a dissimilarity measure can be

reformulated using a notation that uses n_1, n_2, p_1 and p_2, and then see if these terms n_1 and n_2 cancel out in the formula. In other words, we ask whether $D(a, b, c) = D(n_1, n_2, p_1, p_2)$ can be reduced to $D(p_1, p_2)$. The following equivalences are useful for reformulation:

$$a = (a + b) \times (a / (a + b)) = n_1 \times p_1 = (a + c) \times (a / (a + c)) = n_2 \times p_2$$

$$b = (a + b) \times (1 - (a / (a + b))) = n_1 \times (1 - p_1)$$

$$c = (a + c) \times (1 - (a / (a + c))) = n_2 \times (1 - p_2)$$

Using the presence-absence form of each dissimilarity measure (column *Incidence-based* in Table 1), property P10 is verified algebraically by trying to cancel n_1 and n_2 out of the formula.

P10 is a stricter property than *homogeneity* (see P7, second paragraph). It is easy to show that fulfilling P10 leads to a coefficient that is homogeneous (invariant to the total number of species of the data set), because if $D(a, b, c) = D(p_1, p_2)$, and knowing that $p_1 = a / (a + b) = k{\cdot}a / (k{\cdot}a + k{\cdot}b)$ and $p_2 = a / (a + c) = k{\cdot}a / (k{\cdot}a + k{\cdot}c)$, then we have $D(a, b, c) = D(ka, kb, kc)$, which proves *homogeneity*. However, the reverse does not follow. Indeed, a given coefficient may be homogeneous without satisfying P10. An example is the asymmetric binary similarity coefficient proposed by Kulczynski (1928), $S_{12} = a / (b + c)$ (S_{12} in Legendre & Legendre 2012), which is homogeneous but does not fulfil P10.

Proofs for individual coefficients –

(1) **Euclidean distance**. When this distance is calculated on presence-absence data it can be formulated as:

$$E = \sqrt{b + c} = \sqrt{n_1 \cdot (1 - p_1) + n_2 \cdot (1 - p_2)}$$

Because n_1 and n_2 cannot be cancelled out, the Euclidean distance does NOT satisfy P10.

(2) **Manhattan distance**. When this distance is calculated on presence-absence data it can be formulated as:

$$M = |b + c| = |n_1 \cdot (1 - p_1) + n_2 \cdot (1 - p_2)|$$

As before, the Manhattan distance does NOT satisfy P10.

(3) **Jaccard** similarity index (proof thanks to A. Chao).

$$J = \frac{a}{a+b+c} = \frac{a}{2a+b+c-a} = \frac{n_1 p_1}{(n_1 p_1 + n_2 p_2) + n_1(1-p_1) + n_2(1-p_2) - n_1 p_1}$$

$$= \frac{n_1 p_1}{n_1 + n_2 - n_1 p_1} = \frac{1}{(1/p_1) + (1/p_2) - 1}$$

Thus, the Jaccard index DOES satisfy P10. All resemblance coefficients that are equal to the Jaccard index for presence-absence data satisfy P10: the **modified mean character difference**, **coefficient of divergence**, **Canberra metric**, **Wishart coefficient** (1- Similarity ratio), and **abundance-based Jaccard**.

(4) **Sørensen** similarity index (proof thanks to A. Chao) The Sørensen similarity index is

$$S = \frac{2a}{2a+b+c} = \frac{2n_1 p_1}{n_1 p_1 + n_2 p_2 + n_1(1-p_1) + n_2(1-p_2)}$$

$$= \frac{2n_1 p_1}{n_1 + n_2} = \frac{2}{(1/p_1) + [n_2/(n_1 p_1)]} \text{ and, since } n_1 p_1 = n_2 p_2,$$

$$= \frac{2}{(1/p_1) + (1/p_2)}$$

Thus, the Sørensen index DOES satisfy P10. All resemblance coefficients that are equal to the Sørensen index for presence-absence data satisfy P10: the **percentage difference** and **abundance-based Sørensen**.

(5) **Ochiai** similarity index:

$$O = \frac{a}{\sqrt{(a+b)(a+c)}} = \sqrt{\frac{a}{(a+b)} \cdot \frac{a}{(a+c)}} = \sqrt{p_1 \cdot p_2}$$

Thus, the Ochiai index DOES satisfy P10. All resemblance coefficients that are equal to the Ochiai index for presence-absence data satisfy P10: the **Hellinger distance**, **chord distance**, and **abundance-based Ochiai**. In this coefficient, the similarity is the geometric mean of p_1 and p_2.

(6) **Species profile distance**. When this distance coefficient is calculated on presence-absence data, it can be written using the *a*-*b*-*c*-*d* notation as:

$$SP = \sqrt{\frac{b+c}{(a+b)(a+c)}}$$

After reformulating this index, we can reduce it to:

$$SP = \sqrt{\frac{n_1 \cdot (1-p_1) + n_2 \cdot (1-p_2)}{n_1 \cdot n_2}} = \sqrt{\frac{1-p_2}{n_1} + \frac{1-p_1}{n_2}}$$

Because n_1 and n_2 cannot be cancelled out, the species profile distance does NOT satisfy P10.

(7) **Whittaker's index of association**. When this distance coefficient is calculated on presence-absence data, it can be written using the *a*-*b*-*c*-*d* notation as:

$$W = \frac{1}{2}\left(\frac{b}{a+b} + \frac{c}{a+c} + \left|\frac{a}{a+b} - \frac{a}{a+c}\right|\right)$$

After reformulating this index, we can reduce it to:

$$W = \frac{1}{2}\left(\frac{n_1 \cdot (1-p_1)}{n_1} + \frac{n_2 \cdot (1-p_2)}{n_2} + \left|\frac{n_1 \cdot p_1}{n_1} - \frac{n_2 \cdot p_2}{n_2}\right|\right)$$

$$W = \frac{1}{2}\left((1-p_1) + (1-p_2) + |p_1 - p_2|\right) = \frac{1}{2}\left[2 - (p_1 + p_2) + |p_1 - p_2|\right]$$

Thus, the Whittaker index of association also DOES satisfy P10.

(8) **Kulczynski** similarity index. When this *similarity* is calculated on presence-absence data it can be straightforwardly formulated as:

$$K = \frac{1}{2}\left[\frac{a}{a+b} + \frac{a}{a+c}\right] = \frac{1}{2}\left[p_1 + p_2\right]$$

Thus, the Kulczynski similarity also DOES satisfy P10. In this coefficient, the similarity is the arithmetic mean of p_1 and p_2.

P11 – Invariance to the total abundance in each sampling unit. Except when researchers only count and identify a fixed number of individuals (which is often the case in plankton or palaeoecological studies), sampling units in the data table are likely to have different total abundances. Some abundance-based dissimilarity indices are only sensitive to relative abundances per site whereas others reflect differences in site total counts. This property was called "density invariance" by Jost *et al.* (2011). It is not the same as property P7 above. One can check property P11 by determining whether a coefficient is altered when the abundances are multiplied by a constant factor that is different for each sampling unit.

P12 – Coefficients with corrections for undersampling. With higher sampling effort, i.e. larger sampling units, rare species, and in particular those that are not found at the two sites under comparison, are more likely to be observed (Chao *et al.* 2006, Cardoso *et al.* 2009). For that reason, dissimilarity coefficients generally underestimate the dissimilarities among sites, the bias decreasing when sampling effort increases. For some binary similarity coefficients, Chao *et al.* (2006) and Jost *et al.* (2011) suggested abundance-based counterparts that incorporate corrections for undersampling bias.

Property class 4: Ordination-related properties. — The remaining properties are not related to the ecological interpretation of a coefficient or the comparability of beta diversity

values. They are, however, useful for ordination and linear modelling of community composition data.

P13 – Euclidean property of D or $\mathbf{D}^{(0.5)}$. A dissimilarity matrix **D** is *Euclidean* if it can be embedded in a Euclidean space of real axes such that the Euclidean distances among points are equal to the dissimilarity values in **D**. For coefficients that are Euclidean, principal coordinate analysis of **D** produces ordinations that are fully represented in Euclidean space (i.e. without negative eigenvalues). Several coefficients have the Euclidean property. Some coefficients that are not Euclidean for **D** become Euclidean after taking the square root of the dissimilarity values (Gower & Legendre 1986); the resulting matrix, which contains values $[D_{hi}^{0.5}]$, is noted $\mathbf{D}^{(0.5)}$. Legendre & Legendre (2012, Tables 7.2 and 7.3) describe the Euclidean properties of 43 commonly-used similarity and dissimilarity coefficients, including several of the coefficients listed in Table 1.

P14 – Emulated by transformation of the raw frequency data followed by Euclidean distance. Legendre & Gallagher (2001) described how some distance coefficients can be obtained by computing the Euclidean distance (eq. 7 in the main paper) after transforming the raw data values in some appropriate way. Four such transformations are described in Appendix S1. Coefficients that can be obtained in that way are interesting because one can obtain BD_{Total} by computing the transformation and then applying eqs 1-3. Moreover, transformed data allow the computation of the beta diversity contributions of individual species through eqs 4a and 4b (SCBD indices) and of sites through eqs 5a and 5b (LCBD indices). One can also use the transformed data directly in linear modelling of community composition data, e.g. by simple (PCA) or canonical (RDA) ordination, *K*-means partitioning, or multivariate regression tree analysis (MRT), because these methods implicitly preserve the Euclidean distance among sites.

In addition to the coefficients obtained by transformation followed by calculation of the Euclidean distance (Appendix S1), the Whittaker index can also be obtained by applying the Manhattan distance to profile transformed data; see section "The dissimilarity coefficients" in the main paper. This produces twice Whittaker's index of association; for that reason, Whittaker's index was dubbed "relativized Manhattan" by Faith *et al.* (1987). The Manhattan distance is, however, not the distance implicit in linear models, so that Whittaker's index of association does not lend itself to linear modelling nor to the calculation of *Species Contributions to Beta Diversity* (SCDB indices) described in the main paper, eq. 4b.

ACKNOWLEDGEMENTS

The idea for example data set 1 used in paragraph P5 was provided by Prof. Anne Chao, Institute of Statistics, National Tsing Hua University, Hsin-Chu, Taiwan. Prof. Chao's example was expanded for the present Appendix.

REFERENCES

Baselga, A. (2010). Partitioning the turnover and nestedness components of beta diversity. *Global Ecol. Biogeogr.*, 19, 134–143.

Faith, D.P., Minchin, P.R. & Belbin, L. (1987). Compositional dissimilarity as a robust measure of ecological distance. *Vegetatio*, 69, 57–68.

Gower, J.C. & Legendre, P. (1986). Metric and Euclidean properties of dissimilarity coefficients. *J. Classif.*, 3, 5–48.

Hajdu, L.J. (1981). Geographical comparison of resemblance measures in phytosociology. *Vegetatio*, 48, 47–59.

306 Jost, L., Chao, A. & Chazdon, R.L. (2011). Compositional similarity and beta diversity. In:
307 *Biological diversity: frontiers in measurement and assessment* [eds Magurran, A. &
308 McGill, B.]. Oxford University Press, Oxford, England, pp. 66–84.
309 Kulczynski, S. (1928). Die Pflanzenassoziationen der Pieninen. *Bull. Int. Acad. Pol. Sci. Lett. Cl.*
310 *Sci. Math. Nat. Ser. B*, Suppl. II (1927), 57–203.
311 Legendre, P. & Gallagher, E.D. (2001). Ecologically meaningful transformations for ordination
312 of species data. *Oecologia*, 129, 271–280.
313 Legendre, P. & Legendre, L. (2012). *Numerical ecology*. 3rd English edition. Elsevier Science
314 BV, Amsterdam.
315 Orlóci, L. (1978). *Multivariate analysis in vegetation research.* 2nd edition. Dr. W. Junk B. V.,
316 The Hague, The Netherlands.
317 Whittaker, R.H. (1972). Evolution and measurement of species diversity. *Taxon*, 21, 213–251.
318 Wright, D.H. & Reeves, J.H. (1992). On the meaning and measurement of nestedness of species
319 assemblages. *Oecologia*, 92, 416-428.