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.