Appendix S2: Notes on the three novel methods for inferring regional biodiversity patterns from fine-scale samples.

This appendix includes detailed formulation and model description, as well as computer code, for the three Hui models presented in the manuscript.

The challenge of drawing valid inferences about multi-scale species richness within a region or other large area based on a representative sample of fine-scale surveys is an important unresolved challenge in macroecology. A number of approaches have been explored to date (see text), but there remain a wide range of potentially productive avenues that have not yet been explored. Here we set out three such novel approaches. The main aspects have been provided in the main text, and we provide here additional notes for the calculation using these three methods. Before running the following models, the study area (of the 32 datasets) was first divided into grids of particular resolution/scale (e.g. 100km^2, 400km^2 and so on). The following models were run for each grid cell based on samples therein. Fig.S1 provides an illustration of the grid systems applied to the dataset.

To reduce computational demand, we only ran the models for five cells with the most number of records (i.e. most intensely sampled cells) for each scale and reported the average estimates for comparison. Due to the limited number of grid cells at extremely large scales, we only reported the average estimates of two most-sampled 40000km^2 cells and, when relevant, estimates of the most sampled 90000km2 cell. The following models also require a reasonable number of samples within the grid cell (say, >10~15) so that a reliable sampling pattern of species occupancy, frequency and turnover emerges. This requirement normally cannot be fulfilled for the WT and ND subsamples for scales <2500km^2 or for the rest for scales <900km^2. As such, estimates for these fine scales were interpolated from second order splines based on estimates from other scales (largely between 2500km^2 and the full extent) and observed values (at 200m^2 for X-only plots and 210m^2 for X+Linear plots).

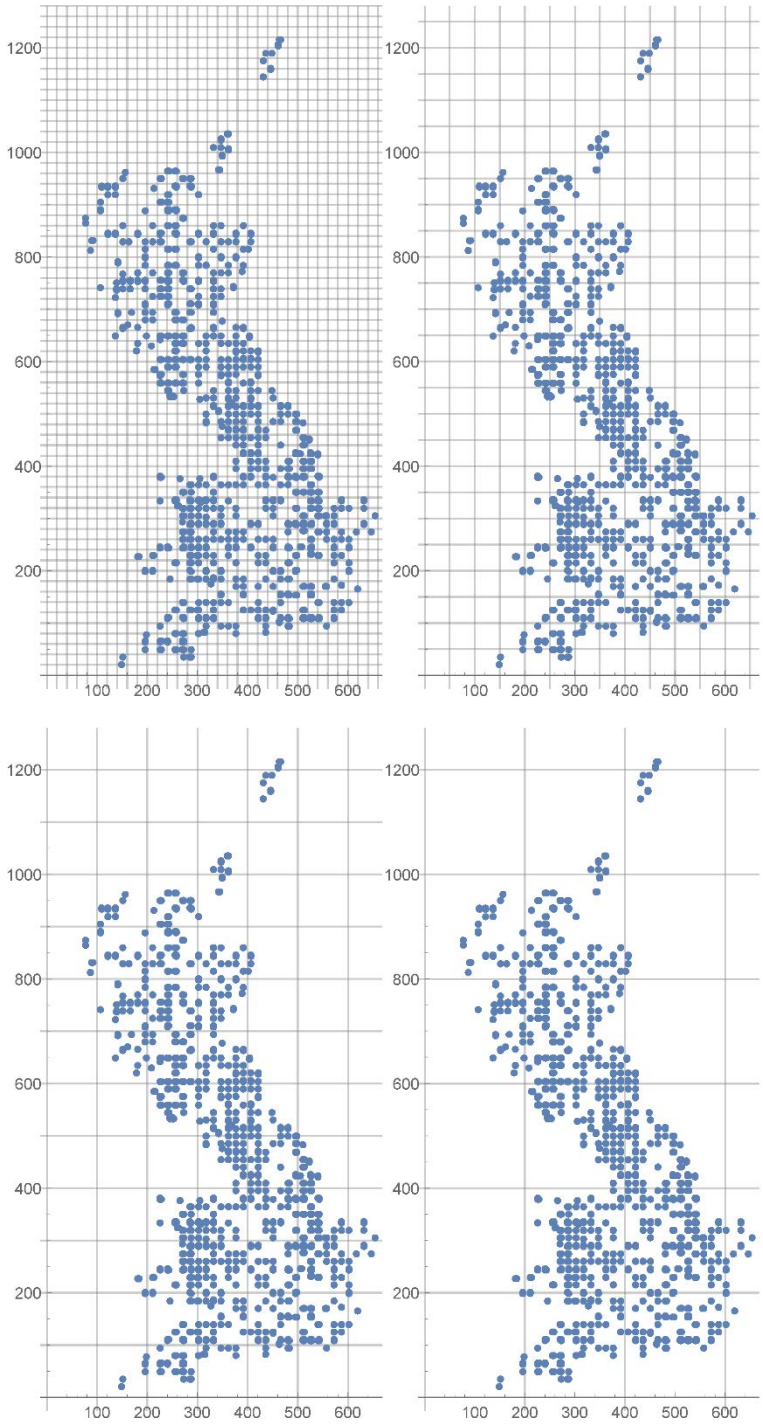

Fig.S1. Examples of grid systems used. From left to right, top to bottom: the grid system at the scale of 20 × 20 km, 50 × 50 km, 100 × 100 km, and 200 × 200 km for the full size X+Linear data.

Hui 1: Occupancy Rank Curve (ORC)

The occupancy rank curve for samples (the number of occupied samples by species rank) generally follows closely a truncated power law (Hui 2012):

$$O = c_1 e^{c_2 \cdot R} R^{c_3},$$

where O and R represent the occupancy and the ranking of a species (R = 1 for the most common species); c_1, c_2 and c_3 are three coefficients. This is the sampling occupancy rank curve (ORC). Such a form of ranked occupancies consists of two components: a power-law function ($c_1 R^{c_3}$) depicting the scale-free structure that no particular scales stand out in the relationship between species ranks and their occupancies, and an exponential cut-off ($e^{c_2 \cdot R}$) depicting a Poission random process of species occupancy. The power-law component is largely applicable to common species, with their distributions reflecting the spatial partitioning (or sharing) of heterogeneous, often fractal, habitat, whilst the exponential cut-off reflects the chance events of the flickering presence/absence of rare species in a homogeneous habitat (or at least perceived as such). The Countryside Survey data fit the truncated power law extremely well (e.g. see Fig.S2).

We begin with a set of n samples with the grain and extent of sampling being a and A, respectively ($A/a = m >> n$; sampling effort = n/m). Assuming that the true and sampling ORCs are of the same shape (i.e. a species with a true occupancy of U at the scale of a having a sampling occupancy of $O = U \cdot n/m$; meaning that the sampling is sufficient and representative), it should be possible to obtain the true ORC by replacing the coefficient c_1 with $C_1 = c_1 \cdot m/n$. The number of species can thus be estimated as the solution for R of the nonlinear equation,

$$1 = C_1 e^{c_2 \cdot R} R^{c_3}.$$

This method essentially blows up the sampling ORC to the true ORC, with the true occupancy then estimated as the sampling occupancy divided by the sampling effort and the maximum ranking for the blown-up ORC thus the true number of species in the sampling extent.

Hui 2: Hypergeometric Discovery Curve (HDC)

Sampling patterns do not necessarily have the same shape as the true macroecological patterns. This is especially true as the probability of discovering a species in a sample does not correlate linearly with species true occupancies. The sampling theory of species abundances that connects true relative abundance distributions to ones emerged from samples has been extensively studied (Dewdney 1998; Green and Plotkin 2007). We here develop a simple method of species occupancies, instead, and its continuation approximation for random sampling. This method is based on assessing how incomplete sampling biases the set of species encountered: the probability of encountering very rare species is near zero, with probability rising with occupancy in a sigmoid fashion and approaching one for very common species.

The probability of discovering a species with a true occupancy of *j* occupying *i* sites amongst a total of *n* samples with the sampling grain *a* over the extent *A* (*m* = *A*/*a*) follows a hypergeometric distribution,

$$prob(i|j) = C_j^i C_{m-j}^{n-i} / C_m^n$$

Non-random sampling or species distributions will obviously complicate the discovery probability, and their effects are ignored here for simplicity. For large *m*, the hypergeometric discovery probability can be approximated by a continuous normal density function $N(i|\mu, \sigma)$ with the mean $\mu = jn/m$ and standard deviation $\sigma = nj(1 - j/m)/m$. We then assess how sampling could affect the shape of observed occupancy frequency distribution (OFD). Let $f(i)$ be the number of species with the sampling occupancy *i* and $F(j)$ the number of species with the true occupancy *j*; that is, the true species richness in an area

$$S = \sum_{j=1}^{m} F(j).$$

As the sampling OFD $f(i)$ is known while the true OFD $F(j)$ unknown, we have the inverse problem of solving the following Fredholm equation of the first kind,

$$f(i) = \sum_{j=1}^{m} prob(i|j)F(j) \approx \int_{j=1}^{m} N(i|\mu, \sigma)F(j)dj.$$

Theoretically, we could assume different parametric forms for the true OFD (e.g., Hui and McGeoch 2007a, b) – a bounded frequency distribution between zero and *m*. In practice, the extremely large number of *m* for this dataset means that we could relax the upper bound and make it simply a nonnegative distribution. One widely-applied nonnegative distribution is lognormal, and for simplicity we thus assume the true OFD follows a lognormal distribution,

$$F(j) = S \cdot LN(j|\mu', \sigma').$$

Species richness S as well as μ' and σ' can be simultaneously determined by minimising

$$\sum_{i=1}^{n} \ln(\hat{f}(i)/f(i))^2,$$

where $\hat{f}(i)$ is the predicted OFD. To substaintly reduce the computational demand, we took the unbiased, symmetric lognormal distribution, with $\mu' = \ln(m)/2$ (the lognormal OFD is centralised around the middle of the possible occupancy at logarithmic scale) and $\sigma' = \ln(m)/3.92$ (the width of the 95% confidence interval spreads the entire possible occupancy at logarithmic scale), making the species richness the sole variable to be estimated from the minimisation.

Hui 3: Zeta diversity

Zeta diversity is a term coined recently to represent the overlap in species across sets of multiple samples (Hui and McGeoch 2014). Unlike pairwise beta diversity which lacks the ability to express the full set of diversity partitions among multiple (≥3) sites, zeta diversity can express the full spectrum of compositional turnover and similarity. Let ζ_j be the number of shared species (intersection) of j randomly selected sites (without replacement) among a total of m sites. In practice, we first fit the zeta diversity decline (i.e. the decline of ζ_j with the increase of zeta order j) to a specific parametric form. As power law and negative exponential are the two most common forms of zeta diversity decline, the use of a truncated power law (exponential power law) will guarantee a good fit. Based on fitted zeta diversity decline, we can estimate the number of species observed in m sites by

$$S_m = \sum_{j=1}^{m} (-1)^{j+1} C_m^j \zeta_j.$$

When m is large, we could use the integral to approximate this (with binomial coefficients replaced by the manipulation of Gamma functions). This allows us to extrapolate zeta diversity with higher orders, and to calculate S_n based on the above formula; notably, it collapses to the Chao II estimator when zeta diversity declines exponentially. When *m* is large, approximation in the above formula often leads to overflowing errors. Instead, we could estimate the number of new species encountered when adding one extra sample (Hui and McGeoch 2014),

$$S_n - S_{n-1} = \frac{\sum_{j=1}^{n} (-1)^{j+1} C_{n-1}^{j-1} \zeta_j}{n} \approx S_{n-1} f_{n-1},$$

where f_n represents the portion of species to be discovered in the extra sample and follows a power law with a negative exponent. That is, we have

$$S_m = S_{m-1}(1 + f_{m-1}) = S_1 \prod_{j=1}^{m-1} (1 + f_j)$$

We estimate the form of f_j based on estimated S_n. Finally, we calculate the integral of $\ln(S_m)$ so that the above iteration can be simplified into the integral over 1 and m. The R implementation of zeta diversity analysis and related multi-site generalised dissimilarity modelling is available in the *zetadiv* package (Latombe et al. 2017a, 2017b).

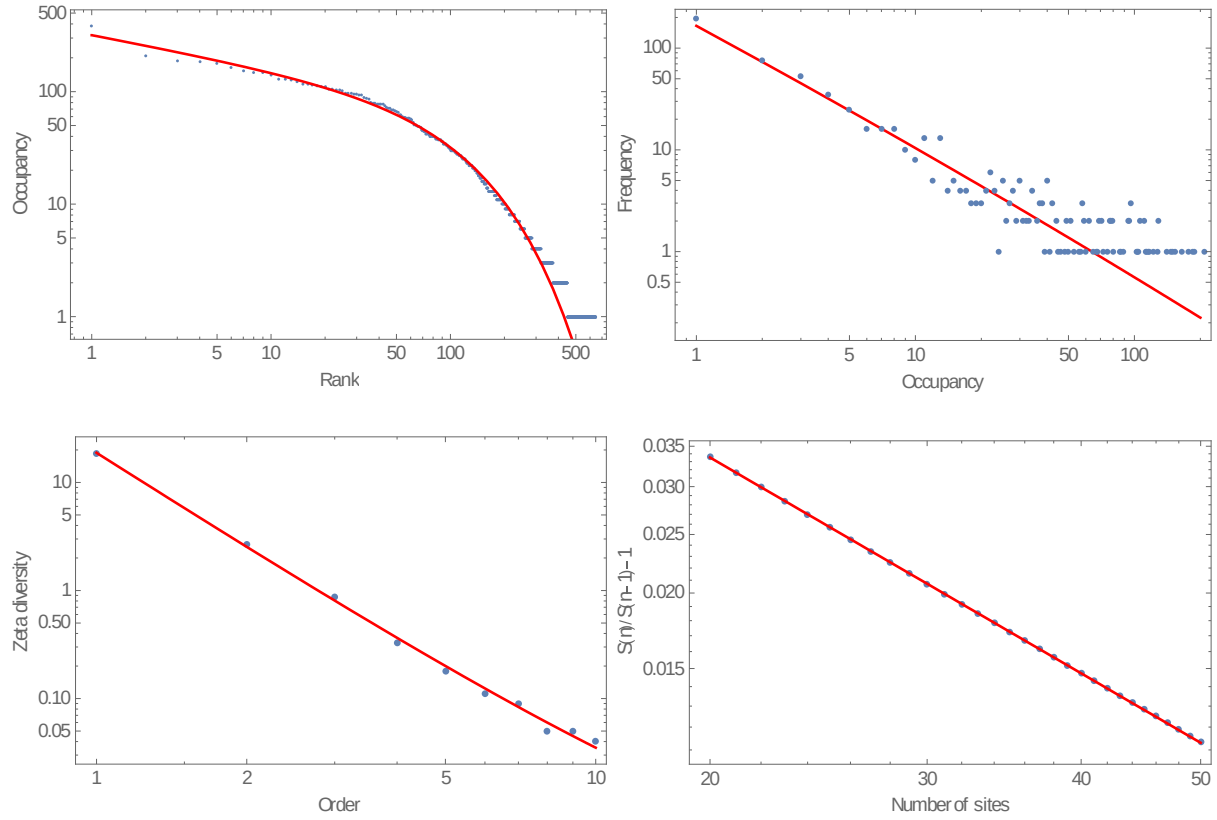

Fig.S2. An illustration of key figures when using the three Hui models for the X-Only WT1 dataset for the full Britain extent. Top left: Occupancy-rank curves (dots: observed; red curve: fitted truncated power law). Top right: Occupancy frequency distributions (dots: observed; red curve: OFD for estimated species richness and the specified true lognormal distribution. Bottom left: Zeta diversity declines (dots: observed mean from 100 combinations; red curve: fitted exponential power law). Bottom right: Portion of species discovered in one extra site (dots: observed; red curve: fitted power law).

Computer code

We implemented the models in Wolfram Mathematica 11.0 with annotations in (* *).

```
(*Data preparation*)
a2 = a; (*a is a dataframe of all records located within a focal cell*)
(*headers of each column were included in the first row*)
xm1 = Dimensions[a2][[1]]; (*# records*)
sit = Tally[Table[a2[[i, 5]], {i, 2, xm1}]]; (*5th col: Rep_ID*)
ns = Dimensions[sit][[1]]; (*# sites*)
site = Table[sit[[i, 1]], {i, 1, ns}]; (*site vector*)
b = Tally[Table[a2[[i, 15]], {i, 2, xm1}]]; (*15th col: Spp_ID*)
sp = Dimensions[b][[1]]; (*# species*)

(*Hui 1: Occupancy Rank Curve*)
b2 = Transpose[a2];
c = Drop[Tally[b2[[15]]], 1];
cc = Sort[Table[c[[i, 2]], {i, sp}], Greater];
data = Table[{i, cc[[i]]}, {i, 1, sp}]; (*ORC*)
nlm = NonlinearModelFit[data, c1 Exp[-c2 z] z^c3, {c1, c2, c3}, z,
  Weights -> Range[Dimensions[c][[1]]]];
Flatten[NSolve[(nmax/ns)*nlm[z] == 1, z]][[1, 2]]; (*# species estimated*)

(*Hui 2: Discovery Curve*)
(*Define Discovery probability*)
cov[i_, j_, n_, m_] :=
 PDF[NormalDistribution[j*n/m, Sqrt[n*j (1 - j/m)/m]], i];
(*Define true OFD*)
ff[j_, u_, v_] := PDF[LogNormalDistribution[u, v], j];
m = 10; (*Only consider the OFD for species with occupancies ≤ m*)
ux = Log[nmax]/2; vx = Log[nmax]/3.92; (*parameters assumed*)
oc = Sort[Table[b[[i, 2]], {i, 1, sp}], Less]; (*Species occupancies*)
ofd = Tally[oc]; (*OFD*)
data = Table[{s,
```

```
   Sum[(Log[
       s NIntegrate[
         cov[i, j, ns, nmax] ff[j, ux, vx], {j, 1, nmax}]] -
       Log[ofd[[i, 2]]])^2, {i, 1,
     Min[Dimensions[ofd][[1]], m]}]}, {s, 100, 5000, 100}]; (*SS for given # species*)
fx = Interpolation[data];
FindMinimum[{fx[x], 100 <= x <= 5000}, {x, 300}][[2, 1, 2]]; (*# species estimated*)
```

(*Hui 3: Zeta Diversity*)

```
Do[{sbs[i, j] = 0}, {i, 1, sp}, {j, 1, ns}];
Do[{sbs[Position[b, a2[[i, 15]]][[1, 1]],
   Position[site, a2[[i, 5]]][[1, 1]]] = 1}, {i, 2, xm1}]; (*Species-by-Site Matrix*)
(*calculating zeta for 100 combinations*)
Do[{
  Do[{sam = RandomSample[Range[ns], k1];
    samm[tt] =
     Total[Table[Product[sbs[i, j], {j, sam}], {i, 1, sp}]]}, {tt, 1, 100}];
  zeta[k1] = Mean[Table[1.0 samm[tt], {tt, 1, 100}]];}, {k1, 1, Min[10, ns]}];
(*Calculating zeta declines using weighted regression*)
nlm = NonlinearModelFit[Table[{k1, zeta[k1]}, {k1, 1, Min[10, ns]}],
  c1 *Exp[-c3*x] x^c2, {c1, c2, c3}, x, Weights -> Range[Min[10, ns]]^4];
(*Calculating # species in n sites*)
Do[{ssm[n] =
   Sum[(-1)^(k1 + 1) Gamma[
     n + 1] nlm[k1]/(Gamma[k1 + 1] Gamma[n - k1 + 1]), {k1, 1,
    n}]}, {n, 1, 100}];
(*Calculating proportion of gained species with one extra sample*)
nlm2 = NonlinearModelFit[
  Table[{n, ssm[n]/ssm[n - 1] - 1}, {n, 20, 50}], c4*x^c5, {c4, c5}, x];
(*Estimated # species*)
Exp[Log[ssm[1]] +
   NIntegrate[Log[1 + nlm2[i]], {i, 1, nmax}, MaxRecursion -> 1000]];
```

Literature Cited

Dewdney, A. K. 1998. A general theory of the sampling process with application to the "veil line". Theoretical Population Biology 54:294–302.

Green, J. L., and J. B. Plotkin. 2007. A statistical theory for sampling species abundances. Ecology Letters 892 10:1037–1045.

Hui, C. 2012. Scale effect and bimodality in the frequency distribution of species occupancy. Community Ecology 13:30–35.

Hui, C., and M. A. McGeoch. 2007a. A self-similarity model for occupancy frequency distribution. Theoretical Population Biology 71:61–70.

Hui, C., and M. A. McGeoch. 2007b. Modelling species distributions by breaking the assumption of self-similarity. Oikos 116:2097–2107.

Hui, C., and M. A. McGeoch. 2014. Zeta diversity as a concept and metric that unifies incidence-based biodiversity patterns. The American Naturalist 184:684–694.

Latombe, G., C. Hui, and M. A. McGeoch. 2017a. Multi-site generalised dissimilarity modelling: Using zeta diversity to differentiate drivers of turnover in rare and widespread species. Methods in Ecology and Evolution 8:431–442.

Latombe, G., M. A. McGeoch, D. A. Nipperess, and C. Hui. 2017b. zetadiv: Functions to compute compositional turnover using zeta diversity. Version 1.0.1, R package. Available at https://CRAN.R-project.org/package=zetadiv