Supplementary Material for: Does systematic heterogenization improve the reproducibility of animal experiments?

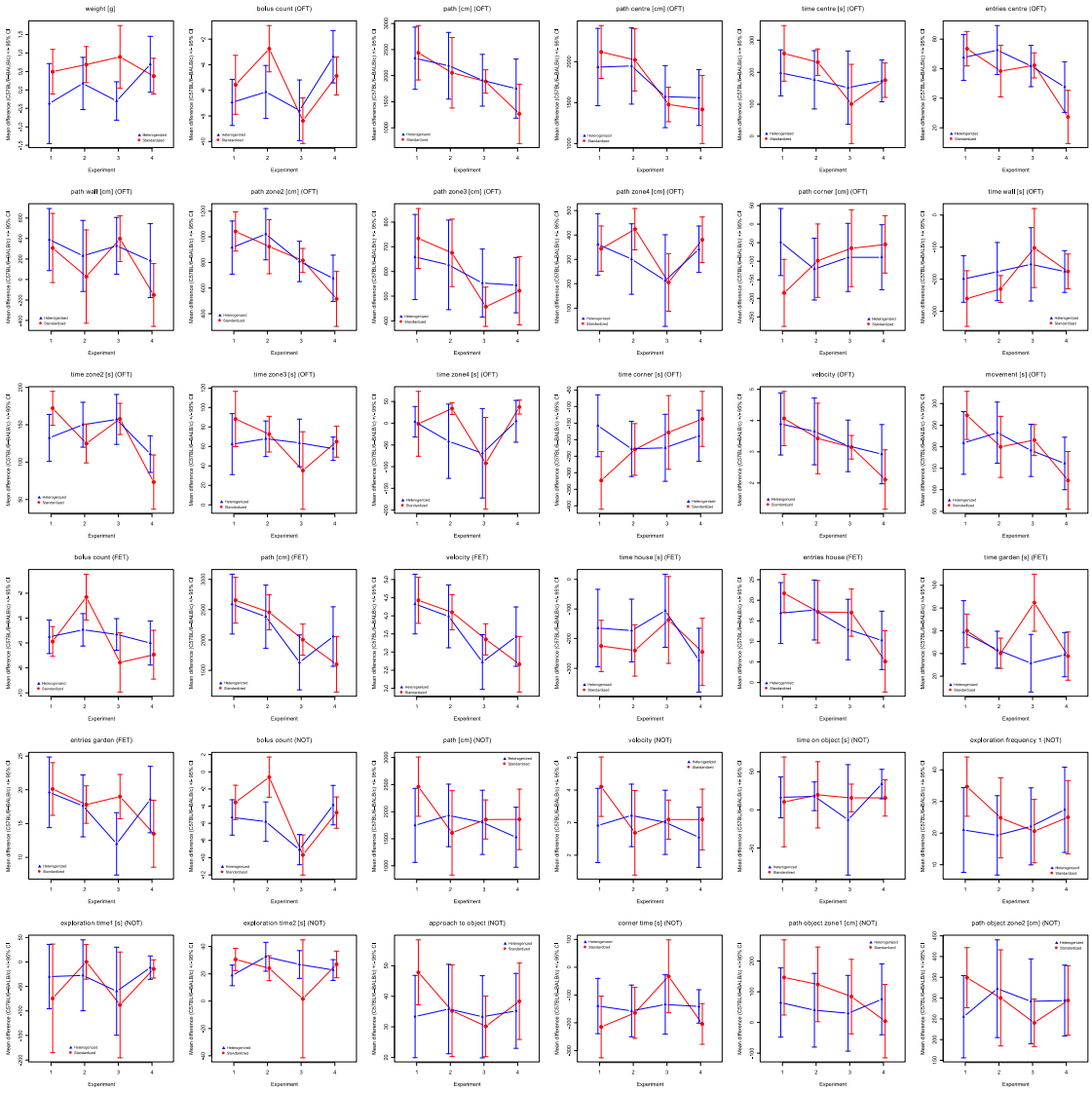
Authors: Rudy M. Jonker[1*], Anja Günther[2], Leif Engqvist[3] & Tim Schmoll[3]

[1]Animal Behaviour, Bielefeld University, Bielefeld, Germany, [2]Behavioural Biology, Bielefeld University, Bielefeld, Germany, [3]Evolutionary Biology, Bielefeld University, Bielefeld, Germany

Supplementary Figure 1	Effects of standardization and heterogenization on between-experiment variation for the total of 36 behavioral measures.
Supplementary Figure 2	Frequency distribution of correlation coefficients between each pair of 36 behavioral measures.
Supplementary Figure 3	Dendrogram for hierarchical clustering of the 36 behavioral measures.
Supplementary Figure 4	Distribution of Pearson correlation coefficients between nine supposedly independent clusters of behavioral measures.
Supplementary Figure 5	Frequency distribution of p-values for the difference between the meta-treatments of all possible models.
Supplementary Figure 6	Variances of behavioral measures for the standardized *versus* heterogenized meta-treatment.
Supplementary Note	Explanation and rationale of analysis.

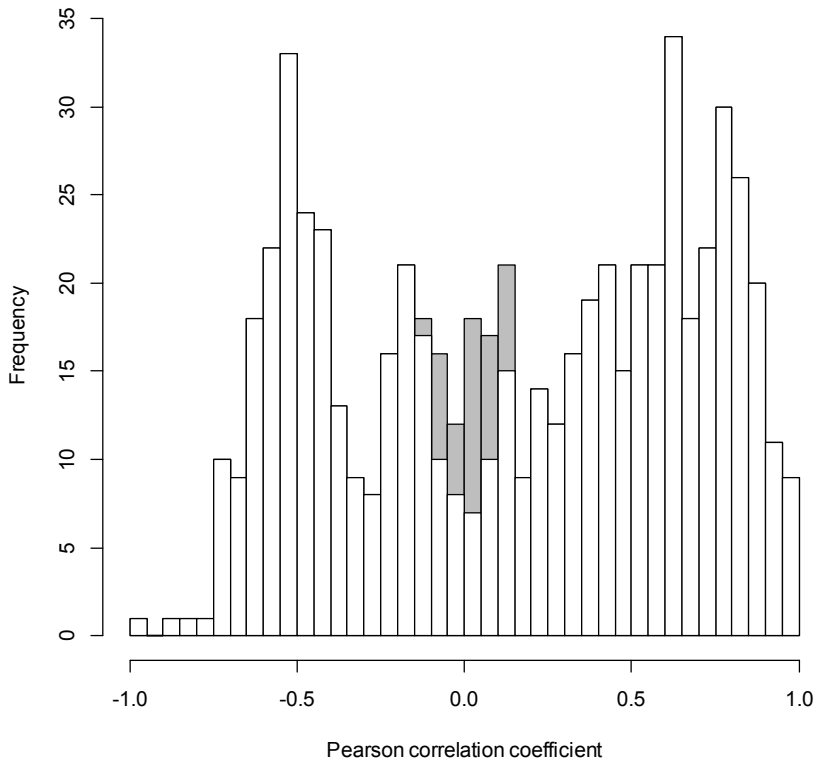
Nature Methods: doi:10.1038/nmeth.2439

Supplementary Figure 1: Effects of standardization and heterogenization on between-experiment variation for the total of 36 behavioral measures. Comparisons of body weight **(1)** and indicated measures from open field (OFT, **2-18**), free exploration (FET, **19-25**) and novel object (NOT, **26-36**) tests show the mean strain differences with their 95% confidence intervals across the four replicate experiments per meta-treatment.



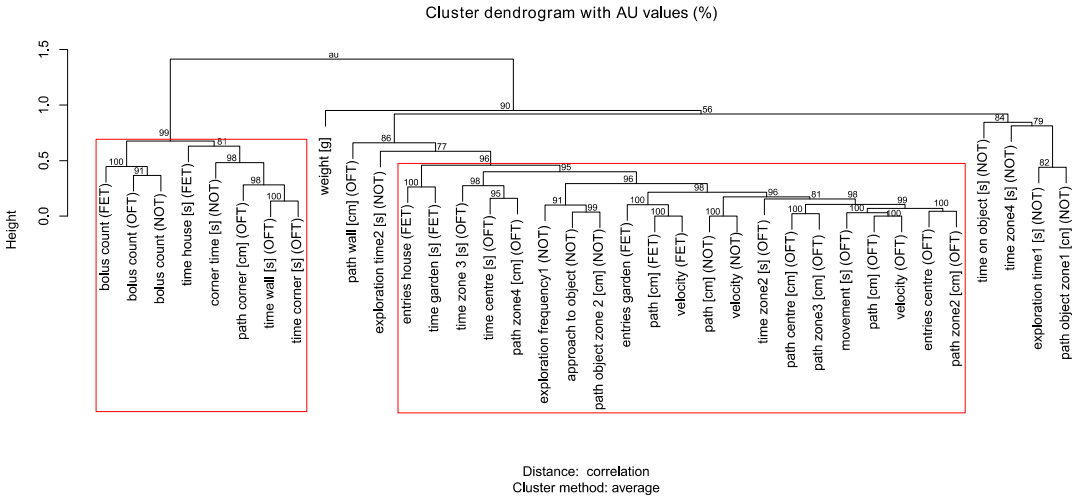
Nature Methods: doi:10.1038/nmeth.2439

Supplementary Figure 2: Frequency distribution of $n = 630$ coefficients (36*35/2) of Pearson's product moment correlations between each pair of 36 behavioral measures. Highlighted in grey are 35 correlation coefficients associated with the response variable body weight, a non-behavioral trait. Note that when using non-parametric Spearman's rank correlations instead, the pronounced bimodal frequency distribution of correlation coefficients becomes even more extreme (data not shown).



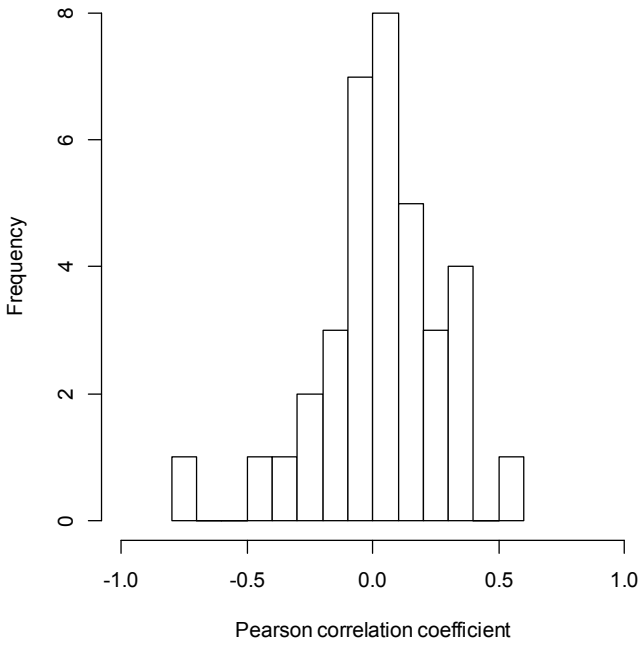
Nature Methods: doi:10.1038/nmeth.2439

Supplementary Figure 3: Dendrogram for hierarchical clustering of the 36 behavioral measures from open field (OFT), free exploration (FET) and novel object (NOT) tests. The *y*-axis shows the dissimilarity *1 – cor(j,k)*. Rectangles indicate groups of variables not separable by multiscale bootstrapping (at $P > 0.05$). AU (approximately unbiased bootstrapping) values indicate levels of concordance in percentage. To assess uncertainty of the clustering, *P* values were calculated using 10000 multiscale bootstraps. Nine clusters were detected at the 0.05 significance level. Variables that were grouped together in one cluster for 8800 out of the 10000 bootstrapping would be given the value AU = 88. Variables that form significant clusters have values above AU = 95.



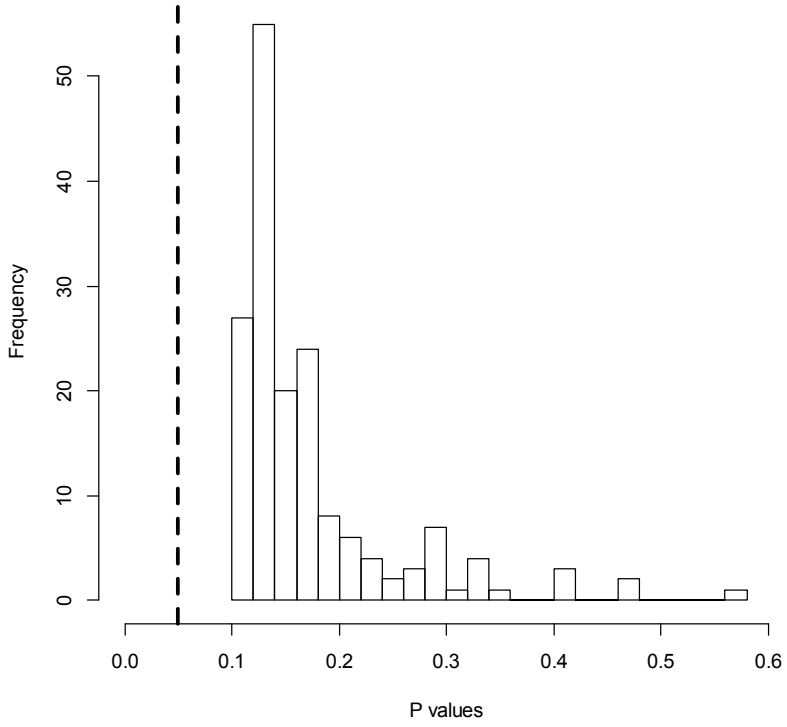
Nature Methods: doi:10.1038/nmeth.2439

Supplementary Figure 4. Frequency distribution of $n = 36$ coefficients of Pearson's product moment correlations between each pair of nine supposedly independent clusters of behavioral measures. Note that many correlation coefficients are around zero, but the relatively long tails of the distribution indicate that there are still some strong correlations between supposedly independent clusters.



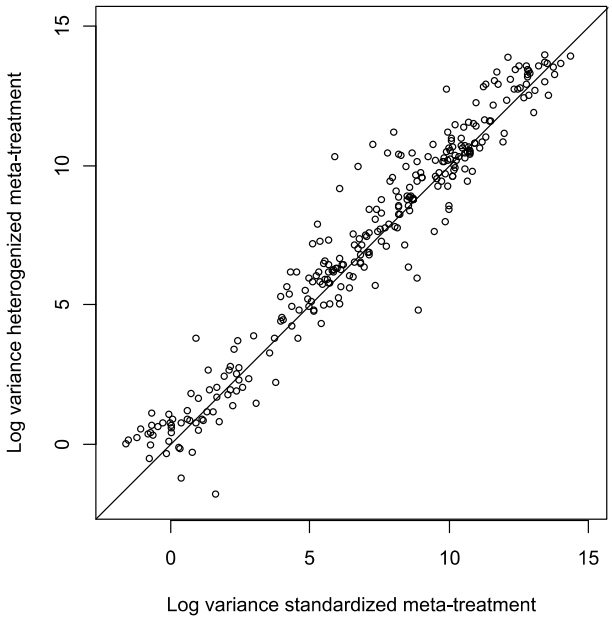
Nature Methods: doi:10.1038/nmeth.2439

Supplementary Figure 5: Frequency distribution of $n = 168$ P values for the significance of a difference between meta-treatments in mean F ratios of the strain-by-experiment interaction terms across all 168 possible combinations of $n = 9$ supposedly independent (clusters of) behavioral measures. The dashed vertical line indicates a significance threshold of $\alpha = 5\%$.



Nature Methods: doi:10.1038/nmeth.2439

Supplementary Figure 6: Log-transformed variances in behavioral measures under a heterogenized *versus* standardized meta-treatment for $n = 288$ pairwise comparisons (variances were calculated separately for the meta-treatments for two strains by four replicate experiments for 36 behavioral measures).



Nature Methods: doi:10.1038/nmeth.2439

Supplementary Note

Error bars

In their Fig. 1 (a-c) Richter *et al*.[1] show mean strain differences across replicate experiments (which intended to simulate different laboratories, but were conducted in the same laboratory) for the heterogenized *versus* standardized experimental design (hereafter meta-treatments), respectively. However, these figures lack information on the confidence of the presented estimates which precludes inference from visual inspection. To quantify the uncertainty of the difference between strain means for each behavioral measure, we calculated standard errors (SE) for these differences using

$$se_{\bar{a}-\bar{b}} = \sqrt{\frac{\sigma_a^2}{n_a} + \frac{\sigma_b^2}{n_b}},$$

where *a* and *b* stand for the different strains, *n* is the sample size and σ^2 is the variance of each strain within each replicate experiment ($n = 16$). To visualize uncertainty, 95% confidence intervals (CI) were calculated as 1.96 · SE. We used 95% CI to allow readers to evaluate possible overlapping with zero and thus to assess whether there was a significant difference in means between the strains (the effect tested for). Adding 95% CI for the mean strain differences discloses an extensive overlap of CI between the replicate experiments across the two meta-treatments for most of the 36 behavioral measures. Thus taking the uncertainty of the differences between the means of replicate experiments into account suggests that there is no conspicuous difference in effect size consistency between the meta-treatments (**Supplementary Fig. 1)**.

Hierarchical clustering

With the notable exception of body weight, a non-behavioral trait, most of the 36 behavioral measures treated as independent by Richter *et al*.[1] are in fact strongly intercorrelated (**Supplementary Fig. 2**). To assess how severely the obvious dependency of behavioural measures may affect the conclusions in Richter *et al*.[1], we identified reasonably independent

Nature Methods: doi:10.1038/nmeth.2439

(groups of) behavioral measures from the pool of 36 using a hierarchical clustering[9] method implemented in the R package *pvclust*[8]. Hierarchical clustering aims at finding groups of samples/variables (behavioral measures here) such that variables within a group are more similar to each other than to variables in different groups. A triangular data matrix consisting of dissimilarities between pairs of variables is the starting point for these analyses. We used hierarchical agglomerative clustering based on group averages. Variables are successively fused into groups and groups into larger clusters, starting with the lowest mutual dissimilarity between variables/groups and then gradually increasing the dissimilarity level at which groups are formed[9]. As dissimilarity measure we used

$$1 - cor(j,k)$$

where $cor(j,k)$ denotes the Pearson correlation between variables j and k. Thus, the correlation coefficients between pairs of variables are transformed into positive dissimilarity values ranging from 0 – 2, which is a stable transformation of correlation coefficients.

To assess uncertainty of the clustering, *P* values were calculated using 10000 multiscale bootstraps. For example, the pairwise correlation of the variables *path [cm] (FET)* and *velocity (FET)* is 1, thus their respective dissimilarity measure is 0 and they are fused together in a cluster in all 10000 bootstraps with high level of concordance (indicated by an approximately unbiased bootstrapping value of AU = 100 in **Supplementary Figure 3**). Successively, the next variable, e.g. *entries of the garden (FET)* is evaluated against the average dissimilarity of *path [cm] (FET)* and *velocity (FET)* (which in this case is still 0). Again, the variable *entries of the garden (FET)* correlates very strongly with the other two variables and are therefore being combined into one cluster in all bootstraps. The average dissimilarity of these three variables is then evaluated against the next variable or the average dissimilarity of a cluster of variables that were formed in the same way (in this case all variables in the large rectangle on the right side of the three abovementioned variables in **Supplementary Figure 3**). These two sub-clusters are then fused together in a bigger cluster

Nature Methods: doi:10.1038/nmeth.2439

because in 98 % of all bootstraps these two sub-clusters cannot be separated from each other (their variables can be in one or the other sub-cluster). Sub-clusters are being fused repeatedly until sub-clusters are reached that can be separated in more than 5 % of the bootstraps (indicated by AU values lower than 95). Applying this approach resulted in nine clusters at the 0.05 significance level, seven of which contained only a single variable and two contained multiple variables (**Supplementary Figure 3** with respective levels of concordance given above the nodes).

To confirm that hierarchical clustering resulted in supposedly independent clusters, we used a Pearson correlation to quantify the still remaining dependencies between each pair of clusters. From the two clusters that contained multiple variables, we selected a variable from the free exploration test as there were no variables from this test in the other clusters (see **Supplementary Figure 3)**. While the distribution of correlation coefficients is now uni- (**Supplementary Fig. 4)** instead of bimodal (cf. **Supplementary Figure 2)**, there are still some relatively strong pairwise correlations suggesting that some of the clusters are in fact not completely independent. We use these nine clusters for subsequent analyses but emphasize that this approach must not be taken to replace independent (series of) experiments for each dependent variable in future test of the heterogenization hypothesis.

Calculation of F-ratios

Following Richter *et al.*[1], we calculated the *F*-ratio of the *strain-by-experiment* interaction term separately for the meta-treatments for each of the nine (clusters of) behavioral measures using the the GLM: y = strain + experiment + block(experiment) + strain x experiment + strain x block(experiment). Exactly following Richter *et al.*[1], we then compared *F*-ratios of the strain-by-experiment interaction terms between the meta-treatments using General Linear Models (GLM) y = meta-treatment + behavioral measure (see Supplementary Methods in Richter *et al.*[1]). However, we analysed all possible combinations of behavioral measures from the aforementioned clusters (resulting in 168 unique variable compositions). For none of these

Nature Methods: doi:10.1038/nmeth.2439

129 168 combinations did the meta-treament have a significant effect on the mean *F*-ratios of the
130 strain-by-experiment interaction terms (**Supplementary Fig. 5**).

131 A further potential source of dependency and hence pseudoreplication arises from the
132 fact that cage mates (there were four individuals per cage) may not only resemble each other
133 because they share the same genetic background (belong to the same strain), but also because
134 they share the same microenvironment, including the social environment. By ignoring the
135 within cage dependency, the true CI for strain differences in **Supplementary Fig. 1** are
136 underestimated. Likewise, the true *P* values for differences between meta-treatments in the
137 strain-by-experiment *F*-ratios in our re-analysis might be even higher than shown in
138 **Supplementary Fig. 5**. In this study, however, re-analysis accounting for the cage effect is
139 impossible with GLMs. Estimating a strain-by-block variance requires at least two
140 independent samples per strain and block. For the same reason, re-analysis using cage means
141 is impossible. In general, a mixed-model framework may be more suitable for analysing data
142 of such structure[9]. However, with the current experimental set-up the variance estimate for the
143 cage effect would be confounded with the variance estimate of the block effect in the
144 heterogenized meta-treatment, as there is only one cage per block per strain.

145 One possible explanation for the fact that the two meta-treatments did not differ in
146 reproducibility might be that they were ineffective in producing levels of sufficiently different
147 within-experiment variation in the behavioral measures, a prerequisite for heterogenization to
148 improve reproducibility (cf. Fig. 2 in Richter *et al.*[1]). This suggestion is supported by a
149 pairwise comparison of variances for the behavioral measures under the two meta-treatments
150 (**Supplementary Fig. 6)**: If the meta-treatment was effective, we would expect variances in
151 the heterogenized experiments to be on average higher than in the standardized experiments,
152 which is not the case. Two-tailed Paired Wilcoxon signed rank tests per behavioral measure
153 showed that only in three out of 36 behavioral measures there was a significantly different
154 variance between meta-treatments (test statistics not shown).

Nature Methods: doi:10.1038/nmeth.2439

155

156

157

158 **References**

159 1. Richter, S. H., Garner, J. P., Auer, C., Kunert, J. & Würbel, H. Systematic variation
160 improves reproducibility of animal experiments. *Nat Meth* **7,** 167–168 (2010).
161 2. Hurlbert, S. H. Pseudoreplication and the Design of Ecological Field Experiments.
162 *Ecological Monographs* **54,** 187–211 (1984).
163 3. Wolf, M. & Weissing, F. J. Animal personalities: consequences for ecology and
164 evolution. *Trends Ecol. Evol.* **27,** 452–461 (2012).
165 4. Lewejohann, L., Zipser, B. & Sachser, N. 'Personality' in laboratory mice used for
166 biomedical research: a way of understanding variability? *Dev Psychobiol* **53,** 624–630
167 (2011).
168 5. Walker, M. D. & Mason, G. Female C57BL/6 mice show consistent individual
169 differences in spontaneous interaction with environmental enrichment that are predicted
170 by neophobia. *Behavioural Brain Research* **224,** 207–212 (2011).
171 6. Sih, A., Bell, A. & Johnson, J. C. Behavioral syndromes: an ecological and evolutionary
172 overview. *Trends in Ecology & Evolution* **19,** 372–378 (2004).
173 7. Schumann, D. E. W. & Bradley, R. A. The Comparison of the Sensitivities of Similar
174 Experiments: Theory. *The Annals of Mathematical Statistics* **28,** 902–920 (1957).
175 8. Suzuki, R. & Shimodaira, H. Pvclust: an R package for assessing the uncertainty in
176 hierarchical clustering. *Bioinformatics* **22,** 1540 –1542 (2006).
177 9. Zuur, A., Ieno, E. N., Walker, N., Saveliev, A. A. & Smith, G. M. *Mixed Effects Models*
178 *and Extensions in Ecology with R*. (Springer, 2009).

179

Nature Methods: doi:10.1038/nmeth.2439