MOLECULAR ECOLOGY

Molecular Ecology (2015) **24**, 759–770 doi: 10.1111/mec.13064

Phylogenetic skew: an index of community diversity

HUNGYEN CHEN,* KWANG-TSAO SHAO† and HIROHISA KISHINO*

*Graduate School of Agricultural and Life Sciences, The University of Tokyo, 1-1-1 Yayoi, Bunkyo-ku Tokyo 113-8657, Japan,
†Biodiversity Research Center, Academia Sinica, 128 Academia Road, Section 2, Nankang Taipei 11529, Taiwan

Abstract

The distribution of divergence times between member species of a community reflects the pattern of species composition. In this study, we contrast the species composition of a community against the meta-community, which we define as the species composition of a set of target communities. We regard the collection of species that comprise a community as a sample from the set of member species of the meta-community, and interpret the pattern of the community species composition in terms of the type of species sampled from the meta-community. A newly defined effective species sampling proportion explains the amount of the difference between the divergence time distributions of the community and that of the meta-community, assuming random sampling. We propose a new index of phylogenetic skew (PS), as the ratio of the maximum-likelihood estimate of the effective species sampling proportion to the observed sampling proportion. A PS value of 1 is interpreted as random sampling. If the value is >1, the sampling is suspected to be phylogenetically skewed. If it is <1, systematic thinning of species is likely. Unlike other indices, the PS does not depend on species richness as long as the community has more than a few members of a species. Because it is possible to compare partially observed communities, the index may be effectively used in exploratory analysis to detect candidate communities with unique species compositions from a large number of communities.

Keywords: community diversity, distribution of divergence times, effective species sampling proportion, phylogenetic tree, species composition

Received 12 October 2014; revision received 18 December 2014; accepted 30 December 2014

Introduction

Because community phylogenetic structure can provide insights into evolutionary processes shaping ecosystem function (Erwin 1991; Davies & Buckley 2011; Mouquet *et al.* 2012; Bell 2013), phylogenetic diversity has been suggested as an additional component of nature conservation assessment (Vane-Wright *et al.* 1991; Faith 1992; Faith *et al.* 2004; Cadotte *et al.* 2012; Srivastava *et al.* 2012). Ecophylogenetics is rapidly becoming an important subfield of community ecology (Mouquet *et al.* 2012) because of its usefulness in addressing questions related to large-scale community spatial patterns (Knapp *et al.* 2008; Crisp *et al.* 2009; Morlon *et al.* 2011; Brum *et al.* 2012), long-timescale variation (Thuiller *et al.* 2011) and spatiotemporal change (Leprieur *et al.* 2011; Jetz *et al.* 2012). Because phylogenies reflect integrated phenotypic differences among taxa, evolutionary relationships may be related to ecological processes and dynamics (Felsenstein 1985; Harvey & Pagel 1991; Faith 1992).

The biodiversity of a community can be measured by its species richness, evenness and abundance distribution (Magurran 2003), with a number of proposed diversity indices also incorporating phylogenetic information (Winter *et al.* 2013). Indices of phylogenetic diversity incorporate information about evolutionary relationships among member species of a community (Cadotte *et al.* 2008). Two major groups of phylogenetic diversity indices that measure either richness or distinctiveness of communities have long been used. Faith's phylogenetic diversity (Faith's PD; Faith 1992), perhaps the most widely used measure of phylogenetic diversity, is a metric quantifying the phylogenetic richness of a community.

Correspondence: Hirohisa Kishino, Fax: +81 3 5841 5066;
E-mail: kishino@lbm.ab.a.u-tokyo.ac.jp

Faith's PD that measures the shared phylogenetic history among taxa occurring in a sample is calculated as the sum of branch lengths from the roots to the tips for the community. The second group of commonly used indices measure phylogenetic diversity in terms of phylogenetic distinctiveness and rely on averaged branch lengths. Average taxonomic distinctiveness (AvTD; Pienkowski *et al.* 1998) is calculated as the sum of all branch lengths connecting two randomly chosen species averaged across all species representing the mean distance between those two species. Rao's quadratic entropy (Rao's QE; Rao 1982) is mathematically similar to AvTD, but can also account for abundances by weighting mean distances between two randomly chosen species. Other phylogenetic diversity indices that measure distinctiveness of communities using branch lengths have recently been proposed for exploring ecological processes. Mean pairwise distance (MPD; Webb 2000) reflects phylogenetic structuring across the entire phylogenetic tree. Mean nearest taxon distance (MNTD; Webb 2000) reflects the phylogenetic structure of the tips of the tree. Net relatedness index (NRI; Webb 2000) and nearest taxa index (NTI) are calculated as standardized MPD and MNTD, assuming random draws of the same number of species from the same phylogeny pool. These indices of phylogenetic diversity can be regarded as summary statistics of tree topologies and branch lengths.

In this study, we interpret the species composition of a community with reference to the species composition of a set of target communities, which we call a meta-community. While a meta-community is usually defined as a set of communities linked by the dispersal of their organisms, here, it simply represents a hypothetical community that consists of all species in the species pool. For example, the species composition of a local community is characterized by comparing it with the global species composition. Negative environmental and anthropological effects on a community may be investigated by adopting the community in the early period as a reference meta-community. We regard the member species of a community as a sample set of the member species from the meta-community. With this interpretation, it becomes possible to express the pattern of a community's species composition in terms of the type of species sampled. We sometimes use the term 'sample' to emphasize this correspondence, even though the communities are not generated by actually sampling from the meta-community.

We propose a new index of phylogenetic skew (PS). This index may be regarded as a counterpart of the net relatedness index (NRI). NRI averages a number of nodes (on the phylogenetic tree of the meta-community member species) that separate all possible pairs of member species of the community, and standardizes by simulating random sampling. In contrast, we compare the distribution of the divergence times between the member species of the community with that of the meta-community. The latter is expressed by two parameters, the net diversification rate and the apparent speciation rate. Given these parameters, the difference between the two distributions is expressed by a new parameter, the effective species sampling proportion, assuming random species sampling. The skew of species sampling contrasts the effective species sampling proportion to the actual sampling proportion.

Materials and methods

The phylogenetic skew index concept

The phylogenetic tree of the member species of a community is a subtree of the phylogenetic tree of the meta-community member species. For short, we simply call them the community and the meta-community phylogenetic trees. When one species is excluded from a phylogenetic tree, the resultant subtree lacks the corresponding terminal branch and the internal node connecting it with the sister species. As divergence times between a pair of sister species are less on average than the divergence times between other pairs of species, the distribution of a community's divergence times is longer on average than the distribution of the meta-community divergence times. The difference between the two distributions grows, as we exclude more species (Fig. 1).

Here, we define the effective species sampling proportion, ρ_E. It explains the difference between the two distributions, assuming that the species composition of a community is obtained by random sampling from the meta-community. If the community consists of closely related species in the meta-community tree, the ρ_E value is larger than the actual species sampling proportion, ρ_O, the ratio of the number of member species from the community to that of the meta-community. However, if the community consists of diverged species in the meta-community tree, ρ_E may be smaller than ρ_O. Therefore, we define the ratio, $PS = \rho_E/\rho_O$, as an index of phylogenetic skew of a community. A PS value of 1 is regarded as a reference value corresponding to random species sampling. If the value is >1, species sampling is likely to be phylogenetically skewed. If it is <1, systematic thinning of species is suspected.

Application to fish impingement sampling

We demonstrate the concept of the phylogenetic skew index through analysis of a fish community monitored monthly at nuclear power plants in northern Taiwan (Liao *et al.* 2004; Shao 2013). The complicated seafloor

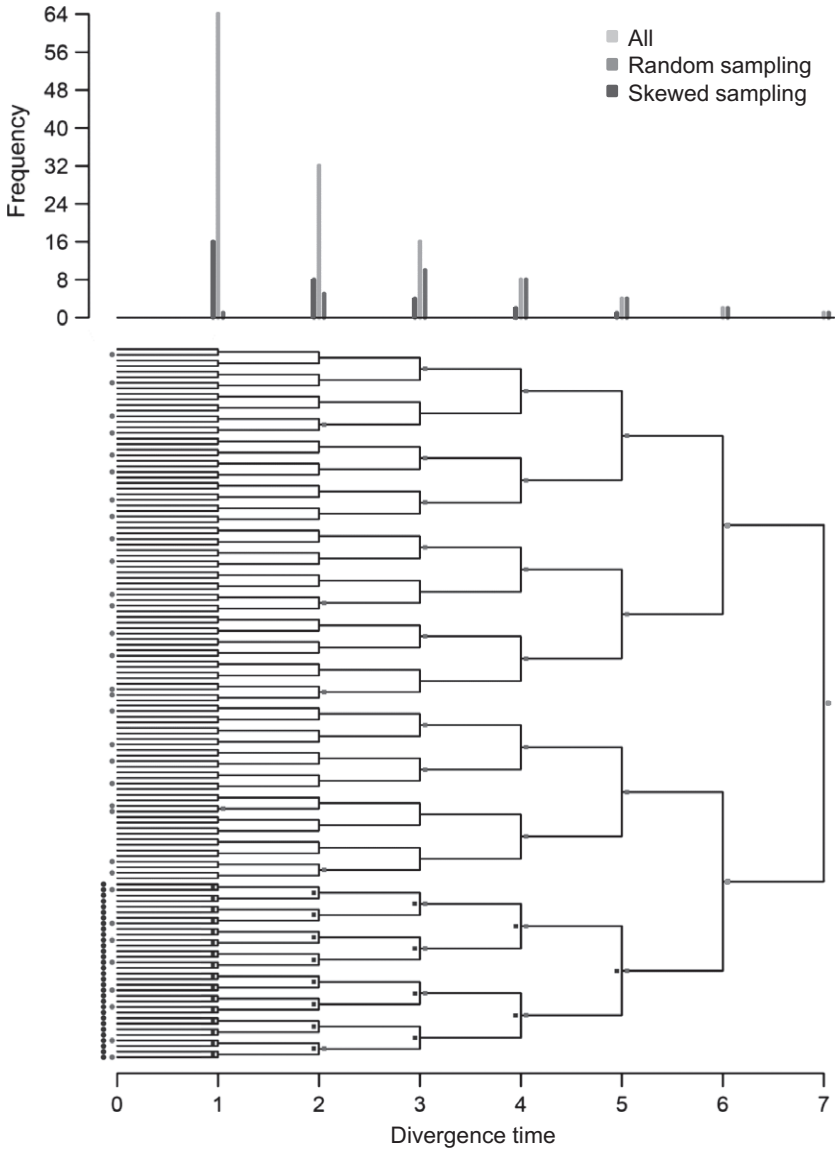

Fig. 1 Divergence time distributions between community members reflect species sampling patterns. Dark grey circles represent random sampling, whereas light grey circles correspond to skewed sampling. Squares mark divergence times. The histogram shows distributions of all divergence times (species sampling proportion = 1), divergence times under random sampling, and divergence times under skewed sampling.

topography, geology and substratum type (e.g. sand, mud, gravel, rock and coral reef) of the surrounding offshore areas, combined with the warm Kuroshio Current, have given rise to a rich and highly diverse aquatic biota. Because of this high species diversity, studies of fish communities adjacent to Taiwan can contribute to many leading global fish diversity indices. Power plant intake screens can serve as ideal locations to monitor fish populations: when large volumes of water are drawn into the plant for cooling purposes, fish are impinged along with the incoming current. Because of the constant rate of water intake, impingement surveys provide informative data for detecting community variation (Greenwood 2008) and can be carried out over a long-time period. The consistency of the collecting scheme combined with fixed sampling efforts decreases systematic errors among samples.

The fish community data used in this study were collected from nuclear power plants at Shihman and Yehliu, both situated in northern Taiwan. Fish samples were collected monthly from intake screens at both plants from July 1987 to April 1990, and from September 2000 to August 2012 (except for December 2006, December 2007 and March 2012 at both plants, and January 2007 at the first plant). Impinged fish were collected from cooling water intakes once every 30 days for 24 h beginning at 9 a.m. on the chosen date using a systematic sampling method (Cochran 1977). The impinged individuals were flushed into a sluiceway and then collected in a trash basket suspended outside the pumping house. All fish were retrieved from the trash baskets and transported to the laboratory for sorting, identification and counting. Because the geographical features of the intakes are similar to one another, we

pooled the monthly samples from both. The increased number of data points enhanced the quality of the pooled data, allowing for more informative analyses. Two hundred and sixty-five species were collected in the 34 months prior to 1990, and 337 species were collected in the 142 months after 2000. The resulting meta-sample comprised 457 species (Table S1; Shao 2013) recorded from the impingement samples.

Bayesian inference of divergence times and phylogenetic skew

Nucleotide sequences of the mitochondrial cytochrome oxidase subunit I (COI) gene were available from the National Center for Biotechnology Information (NCBI 2012) database for 226 species from 73 families of the above-sampled 457 species belonging to 85 families (Table S1, Supporting information). As the coverage of this subsample represents a reasonable proportion of families surveyed, examination of the effect of subsampling on the value of the phylogenetic skew was left for future study. After performing sequence alignments with the MUSCLE program (Edgar 2004) in MEGA 5.0 (Tamura *et al.* 2011), divergence times among the species sampled in the meta-sample were estimated in a Bayesian framework using BEAST v1.7.5 (Drummond *et al.* 2012). The HKY model of nucleotide substitution with gamma-distributed rate heterogeneity among sites (Felsenstein 1981; Hasegawa *et al.* 1985; Yang 1994) was used for the analysis. Because the rate of molecular evolution varies over time, estimation of divergence times based on a molecular clock assumption may be biased (Sanderson 1997; Thorne *et al.* 1998). A random local clock model was used to take variable evolutionary rates among lineages into account (Douzery *et al.* 2002; Drummond & Suchard 2010). The Yule process was used as the prior of the tree. As for the prior distributions of the parameters that specify the substitution process, we adopted the default values. The prior for the HKY transition/transversion parameter was set to log-normal distribution with a location parameter = 1 and a scale parameter = 1.25. The prior distribution of the shape parameter describing the heterogeneity rate among sites was the exponential distribution with mean = 0.5. The frequency of change in evolutionary rate followed a Poisson distribution with mean = 0.7. We did not have the reference nodes for which the time information is available. Because we only use the relative values of the divergence times in the subsequent step, we set the mean evolutionary rate to 1. The Markov chain Monte Carlo (MCMC) chain length was set to 10 000 000. Because the MCMC runs range over tree topologies and the parameters of evolutionary processes, the resulting Bayesian estimates of divergence times take into account the uncertainty of the topology.

Translation of divergence time distribution into a sampling proportion assuming random sampling and calculation of the phylogenetic skew index

The shape of the divergence time distribution depends on three parameters: speciation rate, λ; extinction rate, μ; and species sampling proportion, ρ. Denoting the ordered divergence times (relative, not absolute) of the sample by $\boldsymbol{t} = (t_1, \ldots, t_{s-1})$, $t_1 > \ldots > t_{s-1}$, and using formulae of the generalized birth and death processes (Kendall 1949; Nee *et al.* 1994; Yang & Rannala 1997), the likelihood of $\boldsymbol{t}$ given t_1 is obtained as

$$L_0(\boldsymbol{t}|\lambda, \mu, \rho) = (s-2)! \prod_{j=2}^{s-1} \frac{\lambda p_1(t_j)}{\nu_{t_1}} \qquad \text{eqn 1}$$

where

$$p_1(t) = \frac{1}{\rho} P(0,t)^2 e^{(\mu-\lambda)t}$$

$$P(0,t) = \frac{\rho(\lambda-\mu)}{\rho\lambda + (\lambda(1-\rho)-\mu)e^{(\mu-\lambda)t}},$$

and

$$\nu_t = 1 - \frac{1}{\rho} P(0,t) e^{(\mu-\lambda)t}.$$

The three parameters, λ, μ and ρ, are not identifiable. Specifically, the function, $g(t|\lambda, \mu, \rho) \equiv \frac{\lambda p_1(t)}{\nu_{t_1}}$, that determines the likelihood (eqn 1) can be written as a function of the net diversification rate, $\theta_1 \equiv \lambda - \mu$, and the apparent speciation rate, $\theta_2 \equiv \lambda\rho$:

$$g(t|\lambda, \mu, \rho) = \frac{\theta_2 \left(\frac{\frac{\theta_1}{\theta_2}}{\left(1-e^{-\theta_1 t}\right) + \frac{\theta_1}{\theta_2} e^{-\theta_1 t}} \right)^2 e^{-\theta_1 t}}{1 - \left(\frac{\frac{\theta_1}{\theta_2}}{\left(1-e^{-\theta_1 t_1}\right) + \frac{\theta_1}{\theta_2} e^{-\theta_1 t_1}} \right) e^{-\theta_1 t_1}}$$

In other words, the degrees of freedom of the model is not three but two. With the reparameterized likelihood function, $L_0(\boldsymbol{t}|\theta_1, \theta_2)$, the joint likelihood of the divergence times of species in the meta-community, $\boldsymbol{t}_{\text{meta-comm}}$, and that of species in the community, $\boldsymbol{t}_{\text{comm}}$, is expressed as

$$L(\boldsymbol{t}_{\text{meta}-\text{comm}}, \boldsymbol{t}_{\text{comm}}|\theta_1, \theta_2, \rho_E) = L_0(\boldsymbol{t}_{\text{meta}-\text{comm}}|\theta_1, \theta_2) \times L_0(\boldsymbol{t}_{\text{comm}}|\theta_1, \theta_2 \times \rho_E) \qquad \text{eqn 2}$$

Here, ρ_E can be termed as the effective species sampling proportion. It explains the difference in divergence time distributions of the community and that of the meta-community, assuming that the species compo-

sition of the community is a random sample from the meta-community. Note that the likelihood of the community is a function of the product of θ_2 and ρ_E. When we compare a set of communities, the eqn (2) becomes

$$L\left(\boldsymbol{t}_{\text{meta}-\text{comm}}, \boldsymbol{t}_{\text{comm}}^{(1)}, \ldots, \boldsymbol{t}_{\text{comm}}^{(K)} | \theta_1, \theta_2, \rho_E^{(1)}, \ldots, \rho_E^{(K)}\right) = L_0(\boldsymbol{t}_{\text{meta}-\text{comm}} | \theta_1, \theta_2) \prod_{k=1}^{K} L_0\left(\boldsymbol{t}_{\text{comm}}^{(k)} | \theta_1, \theta_2 \times \rho_E^{(k)}\right) \quad \text{eqn 3}$$

Instead of maximizing the joint likelihood (eqn 3), we adopted the following two-step procedure. In the first step, we estimated the two parameters, θ_1 and θ_2, by maximizing the likelihood of the meta-community, $L_0(\boldsymbol{t}_{\text{meta}-\text{comm}} | \theta_1, \theta_2)$. In the second step, we treated these estimates as fixed and obtained the maximum-likelihood estimate of the effective species sampling proportion for each of the communities by maximizing the partial likelihood $L_0(\boldsymbol{t}_{\text{comm}}^{(k)} | \hat{\theta}_1, \hat{\theta}_2 \times \rho_E^{(k)})$. Furthermore, the observed species sampling proportion was calculated by dividing the number of species in each community by the number of species in the meta-community. The phylogenetic skew index for a community was then defined as the ratio of the estimated effective species sampling proportion to the observed species sampling proportion. Eqns (2) and (3) are not accurate, although each component of them is. The set of species in a community is not independent of the species pool in the meta-community, but rather is a subsample from the species pool. Noting that the information on θ_1 and θ_2 is mostly included in $\boldsymbol{t}_{\text{meta-comm}}$, the two-step procedure approximates the maximum composite likelihood method (Lindsay 1988). A composite likelihood consists of a valid likelihood of subsets of data and has sound theoretical basis and satisfactory performance (Varin & Vidoni 2005).

Comparing phylogenetic diversity indices under three typical sampling scenarios

The values of the proposed PS index and the existing phylogenetic diversity indices depend on the type of species sampled from the meta-community. To gain an insight into the effect of sampling type, we generated community samples from the meta-community impingement data, using three typical types of species sampling: random, quota (systematic thinning) and cluster sampling. To simulate an evenly sampled community among families, we selected one species from each family sampled, with a total of 73 species used for the analysis (Fig. S1a, Supporting information). As a reference, we also generated a random sample of 73 species (Fig. S1c, Supporting information). To simulate a scenario of phylogenetically skewed sampling, we selected the six most dominant families comprising a total of 77 species (Fig. S1e, Supporting information). For each of the simulated communities, we calculated $PS = \hat{\rho}_E/\rho_O$, NRI, NTI, Faith's PD and AvTD values.

Evaluation of small-sample bias in the PS index estimate and bias correction

Our phylogenetic skew index is based on the maximum (composite)-likelihood inference of the effective species sampling proportion, which is efficient as long as the sample size (species richness, in our context) is large. However, the ratio of estimated to observed species sampling proportions, $PS = \hat{\rho}_E/\rho_O$, may be biased, especially when the species richness is small. The number of impinged fish species in each monthly community was not very large and ranged from 3 to 34. To correct for the small-sample-size bias, we conducted a random sampling simulation. After excluding the excessively small community, we generated a set of 1000 random communities from the meta-community for each number of species from 5 to 34. We then calculated the PS value for each community. We fitted gamma distributions to the simulated PS distribution under the null hypothesis of random sampling. Assuming functional dependence of shape (α) and rate (β) parameters on community size (species richness, SR) according to the equations $\alpha = \alpha_{SR} = a_1 + b_1 \times SR$ and $\beta = \beta_{SR} = a_2 + b_2 \times SR$, respectively, we obtained the maximum-likelihood estimates of these parameters. The bias-corrected phylogenetic skew and its 95% confidence interval were calculated by dividing PS by the median and by the 2.5th and 97.5th percentiles of the 1000 PS values estimated assuming a gamma distribution under the null hypothesis of an expected mean of 1.

Results

Phylogenetic skew, NRI, NTI, Faith's PD and AvTD under three typical sampling scenarios

Figure 2a shows the Bayesian phylogenetic COI tree of the 226 fish species in the meta-sample. Based on the estimated divergence times, net diversification rate and apparent speciation rate were estimated as $\hat{\theta}_1 = 4.48 \pm 0.43$ and $\hat{\theta}_2 = 0.87 \pm 0.15$ ($\pm$SE), respectively. The predicted divergence time distribution provided a satisfactory fit to the observed meta-community divergence time distribution (Fig. 2b).

Table 1 shows the $PS = \hat{\rho}_E/\rho_O$, NRI, NTI, Faith's PD and AvTD values for the simulated communities under the three different scenarios. In the case of quota sampling, where one species was sampled from each of the 73 families, the divergence time distribution was longer

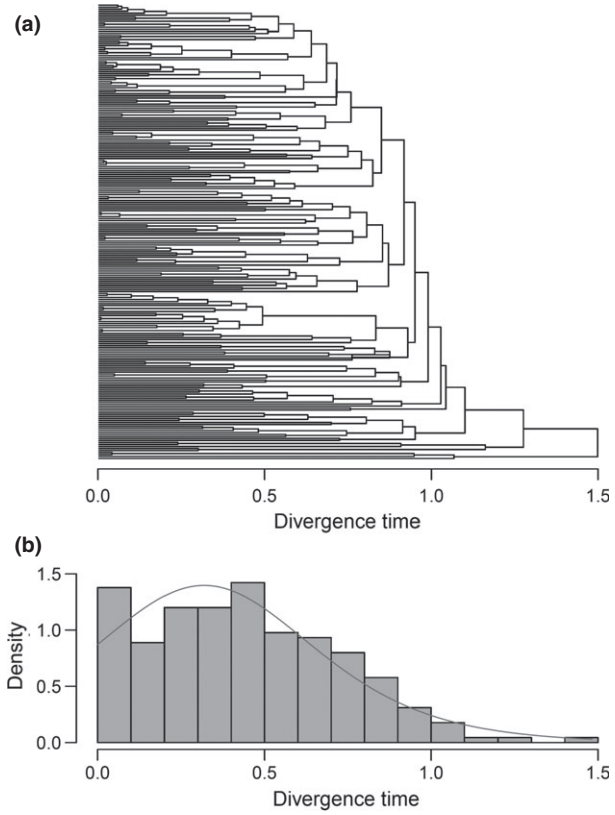

Fig. 2 Bayesian inference of divergence times. (a) Bayesian phylogenetic tree of 226 fish species estimated from COI nucleotide sequences. (b) Histogram of divergence times in the Bayesian phylogenetic tree. The curve represents the distribution of divergence times using eqn (1), with the estimated parameters $\hat{\theta}_1 = 4.47$ and $\hat{\theta}_2 = 0.87$.

on average and the PS was 0.64, far <1. The PS value was close to 1 in the case of random sampling of 73 species. In the case of cluster sampling of 73 species, the divergence time distribution shifted towards the present compared with the distribution obtained by random species sampling.

The Faith's PD values were 54.00, 47.02 and 25.90 for quota, random and cluster sampling, respectively. The AvTD values were 0.1979, 0.1987 and 0.1896 for quota, random and cluster sampling, respectively. Like Faith's PD, the smallest value was obtained when the species were sampled by cluster sampling. AvTD cannot distinguish the difference between quota and random sampling, but can distinguish the difference between cluster and other types of sampling. The NRI (NTI) values were −2.62 (−5.57), −0.14 (−1.18) and 8.05 (8.73) for the three sampling types. The order of the NRI and NTI values was consistent with that of the PS values. As these indices are standardized with reference to the mean and standard deviation of the random sampling, the value was close to zero in the case of random sampling.

Depending on the species sampling scenario, the PS value does not depend on species richness. However, the Faith's PD value depends largely on species richness, and it is difficult to quantify the effect of different types of sampling on phylogenetic diversity. The AvTD value does not depend on species richness. The NRI value decreased when the species was subsampled from the clusters selected in the cluster sampling. Both NRI and PS compare the divergence time distributions among member species with the distribution expected from random sampling. So, it is reasonable to see a similar pattern in the simulation. The difference is the weights on the ancestral nodes. PS puts equal weights on them. However, as a node whose two sister branches have m and n offspring is counted $m \times n$ times, NRI examines the weighted average of divergence times with larger weights on the nodes close to the most recent common ancestor. Like NRI, NTI distinguished the type of sampling well. It still depended on species richness, but was less sensitive than NRI.

Interannual trend and seasonal patterns

Figure 3a shows the PS values of a set of 1000 random communities from the meta-community for each number of species from 5 to 34. The simulated PS distribu-

Table 1 The community diversity indices values for the communities shown in Fig. S1.

	Quota sampling		Random sampling		Cluster sampling		
Species richness	73	50	73	50	77	50	30
PS	0.64	0.71	1.03	1.08	6.28	6.17	6.36
NRI	−2.62	−2.58	−0.14	−0.70	8.05	5.38	3.59
NTI	−5.57	−4.29	−1.18	−1.56	8.73	6.58	5.25
Faith's PD	54.00	39.97	47.02	35.69	25.90	19.91	14.18
AvTD	0.1979	0.1971	0.1987	0.1971	0.1896	0.1908	0.1899

PS, phylogenetic skew; NRI, net relatedness index; NTI, nearest taxa index; Faith's PD, Faith's phylogenetic diversity; AvTD, average taxonomic distinctiveness.

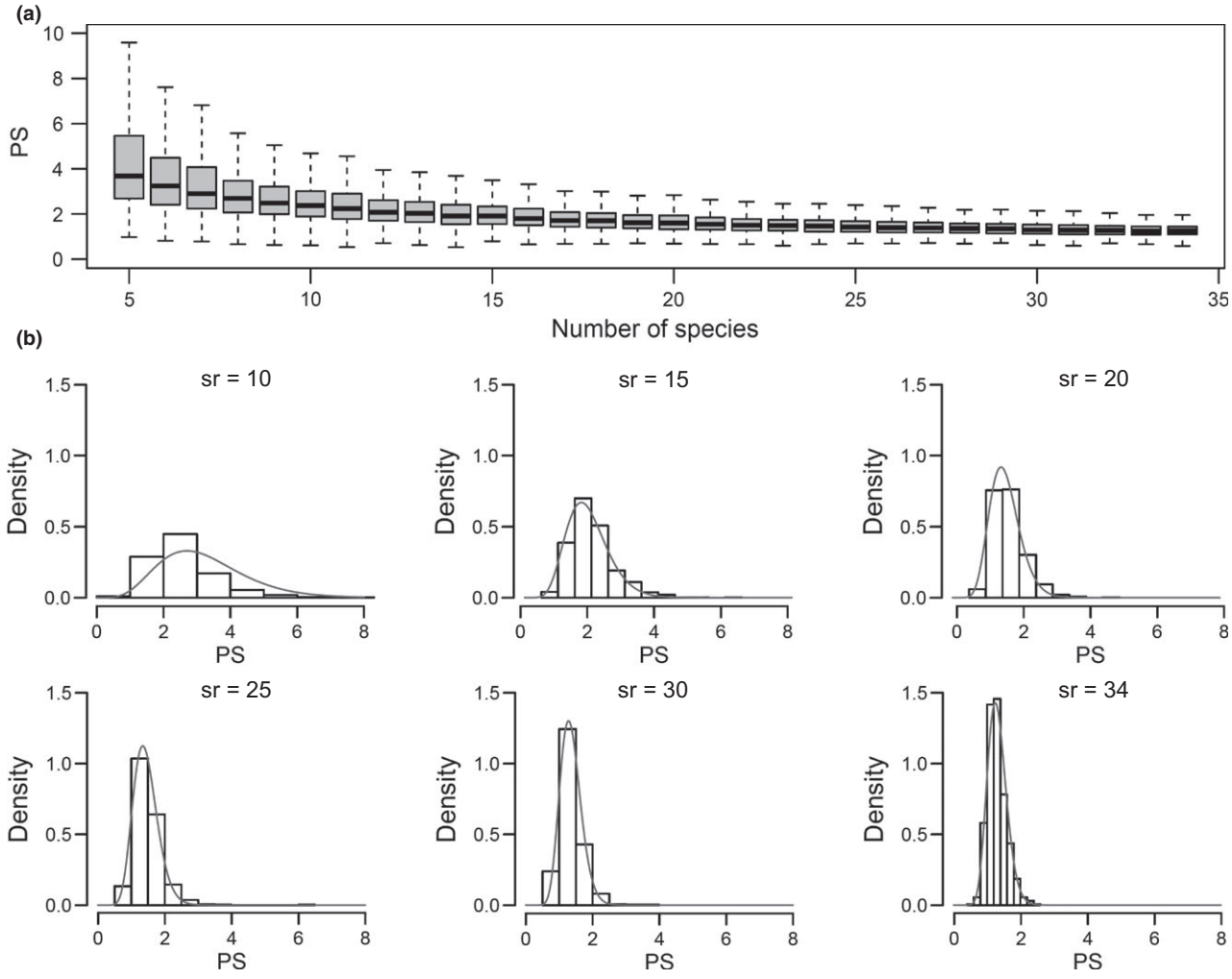

Fig. 3 Phylogenetic skew (PS) calculated from a random sampling simulation. (a) PS box plots of 1000 random samples from the meta-sample for each number of species from 5 to 34. (b) PS histograms for number of species (SR) = 10, 15, 20, 25, 30 and 34 with fitted gamma distributions.

tion under the random sampling null hypothesis was well approximated by a gamma distribution (Fig. 3b). Assuming functional dependence of shape (α) and rate (β) parameters on species richness, SR, to be $\alpha = \alpha_{SR} = a_1 + b_1 \times SR$ and $\beta = \beta_{SR} = a_2 + b_2 \times SR$, respectively, we obtained the maximum-likelihood estimates of the four parameters as $a_1 = 3.49 \pm 0.16$, $b_1 = 0.44 \pm 0.01$, $a_2 = -1.50 \pm 0.08$ and $b_1 = 0.45 \pm 0.01$. Figure 3b shows a satisfactory fit of the model. A plot of the monthly estimated $PS = \hat{\rho}_E/\rho_O$ values from September 2000 to August 2012 compared with that derived from the simulated random sampling distribution is shown in Fig. 4a. The phylogenetic skew and its 95% confidence interval were calculated by dividing PS by the median and by the 2.5th and 97.5th percentiles of the 1000 PS values estimated assuming a gamma distribution under the null hypothesis of an expected mean of 1. The bias-corrected phylogenetic skew and its 95% confidence interval (number of species from 8 to 34) are shown in Fig. 4b. This value ranged from 0.75 to 4.16, with a median of 1.39 and a mean of 1.44 ± 0.19, which is significantly larger than 1 ($P < 0.001$).

A declining trend was observed in the species richness of the impingement fish community. Species richness declined significantly among years (correlation coefficient = −3.98, P-value = 0.002; Fig. 5a) and severely in summer and autumn (Fig. 5b). Figure 5b shows the monthly variation of species richness from July 1987 to April 1990 and from September 2000 to August 2012. The monthly species richness ranged from 3 to 41. The yearly trends were different between species richness and PS (Fig. 5a). Before 2007, although species richness declined, PS did not change. This

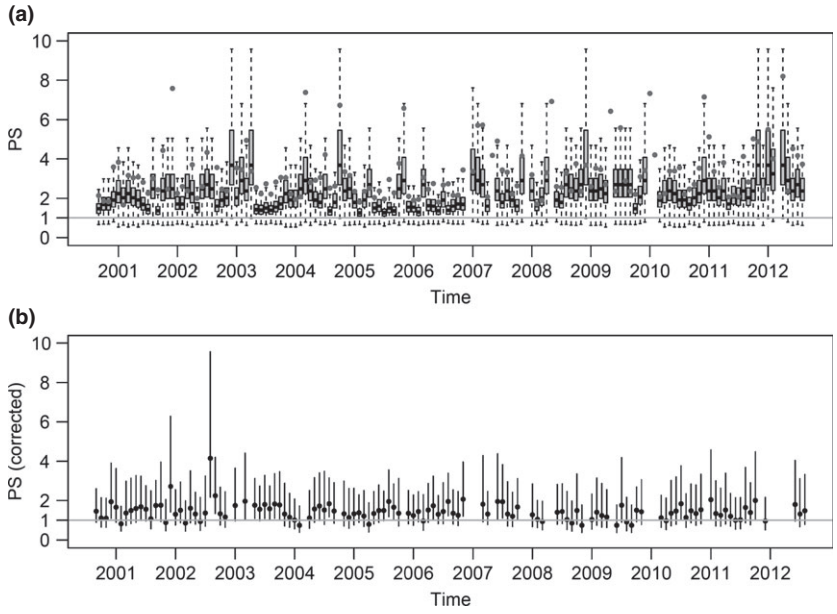

Fig. 4 Time series of monthly phylogenetic skew (PS) of the impingement fish community from September 2000 to August 2012. (a) Observed phylogenetic skew (points) compared with a simulated random sampling distribution (boxes). (b) The bias-corrected phylogenetic skew (points) and its 95% confidence interval (lines).

implies that the species declined systematically and the composition of present species is phylogenetically evenly distributed, more similar to quota sampling. From 2007, PS fluctuated annually when species richness started to decline sharply. The rapid decrease devastated the evenly distributed phylogenetic species composition in the community. The monthly PS (after correction of small-sample bias) from July 1987 to April 1990 and from September 2000 to August 2012 ranged between 0.23 and 4.16. A significant difference between the periods before 1990 and after 2000 is shown in Fig. 5c ($t = -17.29$, P-value <0.001). We then compared the species compositions among the periods 1988–1990, 2001–2003 and 2010–2012 (Fig. 6). An apparent qualitative change in species composition among different periods is shown in the figure. Compared to the 1988–1990 period, species in 2001–2003 decreased more systematically than in 2010–2012. Some species appeared frequently prior to 1990 and disappeared after 2000, for example the species belonged to the families Labridae, Carangidae and Leiognathidae. Leiognathidae and Carangidae are more common in warmer months, suggesting that the cause of their decrease may be related to changing sea temperatures in recent years (Fig. S4, Supporting information).

Discussion

As the PS index is based on a fully parametric model, a natural worry is a possible bias caused by model misspecification. The random speciation and extinction model was used to describe the divergence time distribution between the meta-community member species. The real pattern of speciation and extinction may be far more complex. However, we had only hoped that the model with the net diversification and apparent speciation rate would achieve a satisfactory fit to the observed distribution in most cases. These two parameters are nuisance ones for our purpose. Our target parameter of interest is the effective species sampling proportion, which explains the amount of difference between the divergence time distributions among the member species of a community and that of the meta-community. As the model assumes random sampling, the difference between the estimated parameter value and the actual species sampling proportion is a departure of the community species composition from the pattern expected from random sampling of the meta-community.

The numerical simulation that reflects the fish assemblage in northern Taiwan showed that the maximum-likelihood estimate of the effective sampling proportion is positively biased, when the sample size is small. In this study, we corrected the bias by examining the distribution of the estimate under the random sampling scenario. The simulation showed that the bias decreased monotonically with sample size (species richness). This type of bias may be better handled in the framework of a penalized likelihood or Bayesian approach, where the relative strength of the penalty against the departure from the actual sampling proportion decreases with increasing sample size. The validity of the bias-correction methods need be tested both with extensively designed simulations and empirical data analysis in the future. We wish to propose the PS index for comparative studies of relatively large communities

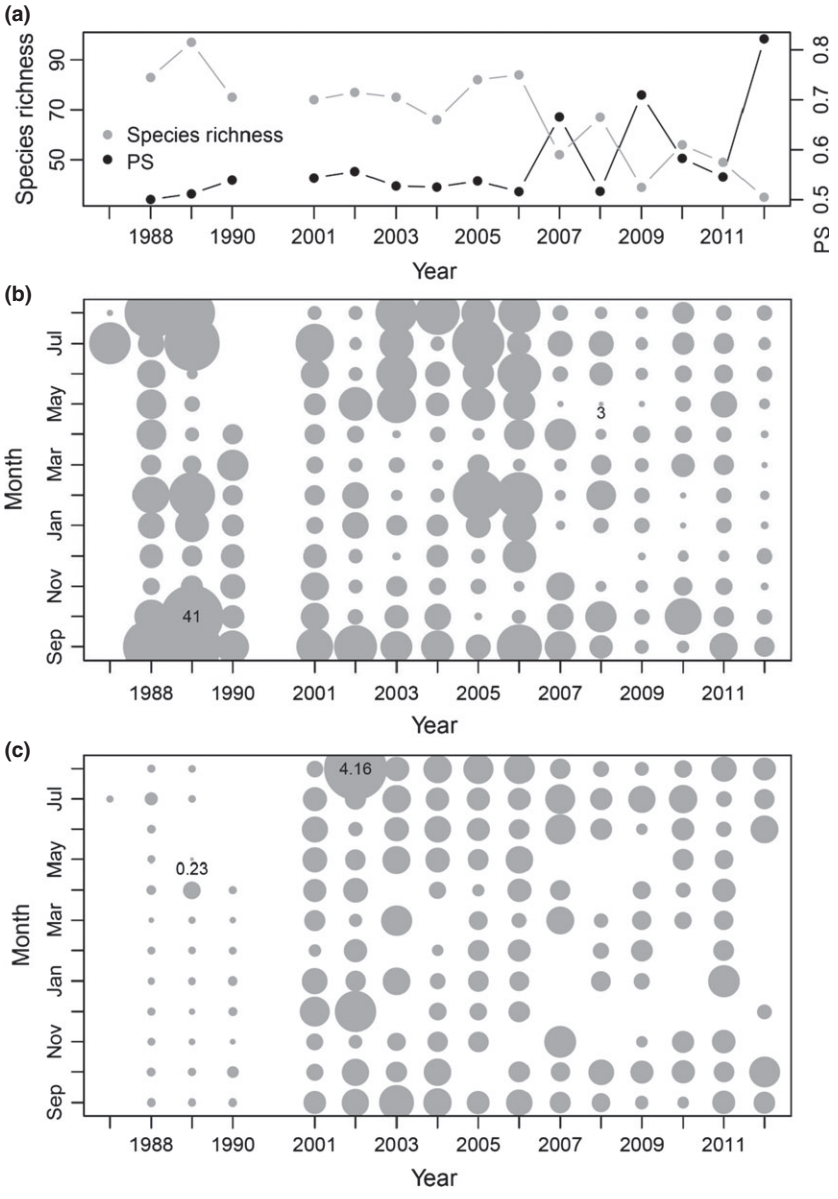

Fig. 5 Temporal variation of species richness and phylogenetic skew (PS) of the impingement fish community. (a) Time series of yearly species richness and phylogenetic skew from 1988 to 1990 and from 2001 to 2012. (b) Time series of monthly species richness from July 1987 to April 1990 and from September 2000 to August 2012. (c) Time series of monthly PS (after the correction of small-sample bias) from July 1987 to April 1990 and from September 2000 to August 2012. Filled circles in (b) and (c) indicate the value in each month. The diameter of the circle is proportional to the value.

that have sufficient information on divergence time distributions. Because the index does not depend on species richness as long as the community has more than a few member species, it enables us to compare the phylogenetic skew between partially observed communities. It may be used effectively in exploratory analysis to detect candidate communities with unique species compositions from a large number of communities. By looking closely at the species compositions of the candidate communities with high phylogenetic skew values, it may be possible to identify factors with unique features.

Remarkable developments in phylogenetic inference procedures and the reduction in sequencing costs have enabled statistical modelling of phylogenetic diversity. The proposed index is computationally demanding. The largest computational burden is the estimation of the divergence times of member species in the metacommunity without assuming a molecular clock. As divergence times depend on tree topology, this method is ideal for taking uncertainty in tree topology inferences into account. However, the divergence time distributions may be relatively robust against the errors in the minute phylogenetic order of the internal nodes that are separated by short branches. The computation becomes inexpensive by adopting the two-stage procedure: the estimation of the branch lengths and the estimation of divergence times from the

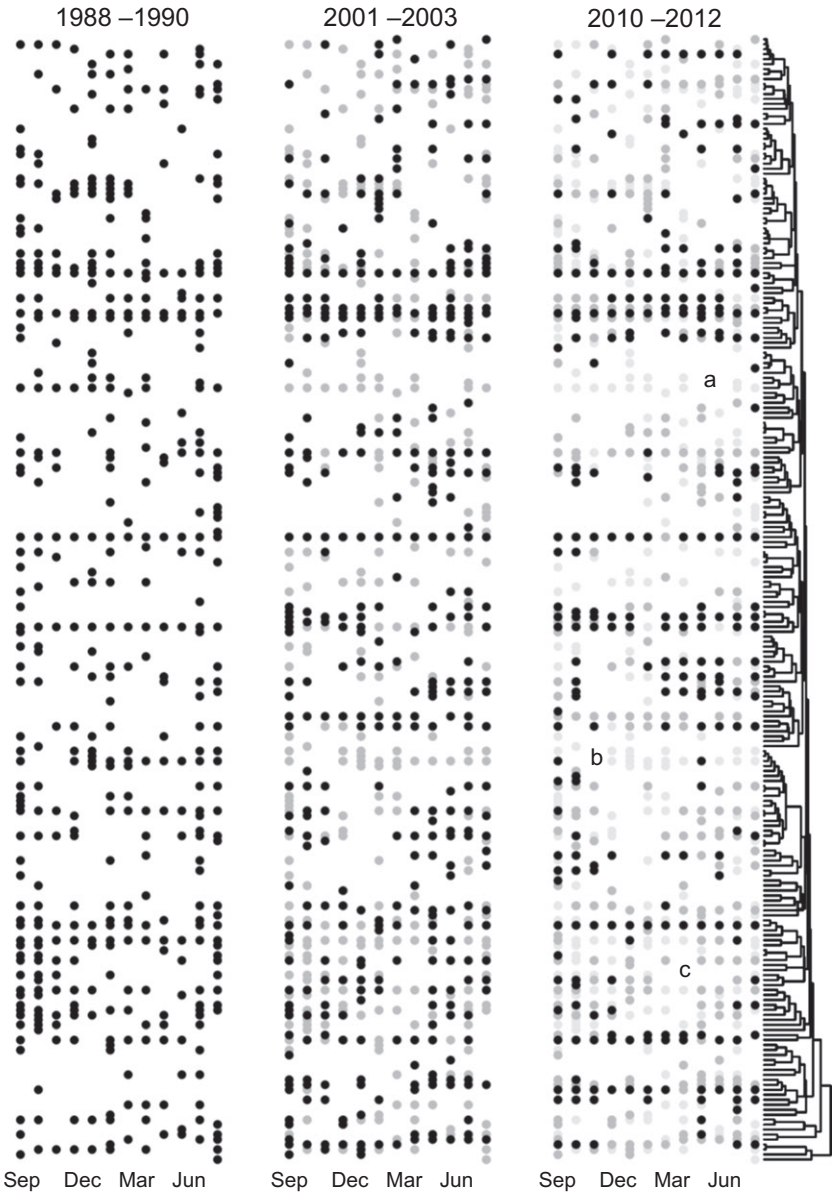

Fig. 6 Species composition of the impingement fish community during the 12-month periods of 1988–1990, 2001–2003 and 2010–2012. The black-filled circle represents the species present. The dark grey-filled circles represent species present in the previous period but absent in the current one. The light grey-filled circle represents the species absent from the last two periods. Family a, Labridae; b, Carangidae; and c, Leiognathidae. See Figs S2 and S4 (Supporting information) for family and species names, respectively.

estimated branch lengths (Thorne *et al.* 1998; Dos Reis & Yang 2011).

Acknowledgements

The authors thank the anonymous reviewers for their valuable comments. This work was supported in part by funding from the Japan Society for the Promotion of Science to HK (grant number 25280006) and funding by Taiwan Power Company for a long-term monitoring project to KS.

References

Bell G (2013) The phylogenetic interpretation of biological surveys. *Oikos*, **122**, 1380–1392.

Brum FT, Kindel A, Hartz SM, Duarte LDS (2012) Spatial and phylogenetic structure drive frugivory in Tyrannidae birds across the range of Brazilian Araucaria forests. *Oikos*, **121**, 899–906.

Cadotte MW, Cardinale BJ, Oakley TH (2008) Evolutionary history and the effect of biodiversity on plant productivity. *PNAS*, **105**, 17012–17017.

Cadotte MW, Dinnage R, Tilman D (2012) Phylogenetic diversity promotes ecosystem stability. *Ecology*, **93**, 223–233.

Cochran WG (1977) *Sampling Techniques*, 3rd edn. John Wiley & Sons, New York.

Crisp MD, Arroyo MTK, Cook LG *et al.* (2009) Phylogenetic biome conservatism on a global scale. *Nature*, **458**, 754–758.

Davies TJ, Buckley LB (2011) Phylogenetic diversity as a window into the evolutionary and biogeographic histories of present-day

richness gradients for mammals. *Philosophical Transactions of the Royal Society Biological Sciences*, **366**, 2414–2425.
Dos Reis M, Yang Z (2011) Approximate likelihood calculation on a phylogeny for Bayesian estimation of divergence times. *Molecular Biology and Evolution*, **28**, 2161–2172.
Douzery EJP, Delsuc F, Stanhope MJ, Huchon D (2002) Local molecular clocks in three nuclear genes: divergence times for rodents and other mammals and incompatibility among fossil calibrations. *Journal of Molecular Evolution*, **57**, S201–S213.
Drummond AJ, Suchard MA (2010) Bayesian random local clocks, or one rate to rule them all. *BMC Biology*, **8**, 114.
Drummond AJ, Suchard MA, Xie D, Rambaut A (2012) Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Molecular Biology and Evolution*, **29**, 1969–1973.
Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, **32**, 1792–1797.
Erwin TL (1991) An evolutionary basis for conservation strategies. *Science*, **253**, 750–752.
Faith DP (1992) Conservation evaluation and phylogenetic diversity. *Biological Conservation*, **61**, 1–10.
Faith DP, Reid CAM, Hunter J (2004) Integrating phylogenetic diversity, complementarity, and endemism for conservation assessment. *Conservation Biology*, **18**, 255–261.
Felsenstein J (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution*, **17**, 368–376.
Felsenstein J (1985) Phylogenies and the comparative method. *The American Naturalist*, **125**, 1–15.
Greenwood MFD (2008) Trawls and cooling-water intakes as estuarine fish sampling tools: comparisons of catch composition, trends in relative abundance, and length selectivity. *Estuarine, Coastal and Shelf Science*, **76**, 121–130.
Harvey PH, Pagel M (1991) *The Comparative Method in Evolutionary Biology*. Oxford University Press, Oxford, UK.
Hasegawa M, Kishino H, Yano T (1985) Dating the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution*, **22**, 160–174.
Jetz W, Thomas GH, Joy JB, Hartmann K, Mooers AO (2012) The global diversity of birds in space and time. *Nature*, **491**, 444–448.
Kendall DG (1949) Stochastic processes and population growth. *Journal of the Royal Statistical Society B*, **11**, 230–264.
Knapp S, Kühn I, Schweiger O, Klotz S (2008) Challenging urban species diversity: contrasting phylogenetic patterns across plant functional groups in Germany. *Ecology Letters*, **11**, 1054–1064.
Leprieur F, Tedesco PA, Hugueny B *et al.* (2011) Partitioning global patterns of freshwater fish beta diversity reveals contrasting signatures of past climate changes. *Ecology Letters*, **14**, 325–334.
Liao YC, Chen LS, Shao KT, Tu YY (2004) Temporal changes in fish assemblage from the impingement data at the second nuclear power plant, northern Taiwan. *Journal of Marine Science and Technology*, **12**, 411–417.
Lindsay BG (1988) Composite likelihood methods. *Contemporary Mathematics*, **80**, 221–239.
Magurran AE (2003) *Measuring Biological Diversity*. Blackwell Science, Oxford, UK.
Morlon H, Schwilk DW, Bryant JA *et al.* (2011) Spatial patterns of phylogenetic diversity. *Ecology Letters*, **14**, 141–149.
Mouquet M, Devictor V, Meynard CN *et al.* (2012) Ecophylogenetics: advances and perspectives. *Biological Reviews*, **87**, 769–785.
NCBI (2012) Nucleotide. Available at: http://www.ncbi.nlm.nih.gov/nuccore. Last accessed 30 March 2012.
Nee S, May RM, Harvey PH (1994) The reconstructed evolutionary process. *Philosophical Transactions of the Royal Society Biological Sciences*, **344**, 305–311.
Pienkowski MW, Watkinson AR, Kerby G, Warwick RM, Clarke KR (1998) Taxonomic distinctness and environmental assessment. *Journal of Applied Ecology*, **35**, 532–543.
Rao CR (1982) Diversity and dissimilarity coefficients—a unified approach. *Theoretical Population Biology*, **21**, 24–43.
Sanderson MJ (1997) A nonparametric approach to estimating divergence times in the absence of rate constancy. *Molecular Biology and Evolution*, **14**, 1218–1231.
Shao KT (2013) The Fish Database of Taiwan, version 2009/1. Available at: http://fishdb.sinica.edu.tw. Last accessed 1 February 2013.
Srivastava DS, Cadotte MW, MacDonald AAM, Marushia RG, Mirotchnick N (2012) Phylogenetic diversity and the functioning of ecosystems. *Ecology Letters*, **15**, 637–648.
Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular Biology and Evolution*, **28**, 2731–2739.
Thorne JL, Kishino H, Painter IS (1998) Estimating the rate of evolution of the rate of molecular evolution. *Molecular Biology and Evolution*, **15**, 1647–1657.
Thuiller W, Lavergne S, Roquet C, Boulangeat I, Lafourcade B, Araujo MB (2011) Consequences of climate change on the tree of life in Europe. *Nature*, **470**, 531–534.
Vane-Wright RI, Humphries CJ, Williams PH (1991) What to protect?—systematics and the agony of choice. *Biological Conservation*, **55**, 235–254.
Varin C, Vidoni P (2005) A note on composite likelihood inference and model selection. *Biometrika*, **92**, 519–528.
Webb CO (2000) Exploring the phylogenetic structure of ecological communities: an example for rain forest trees. *The American Naturalist*, **156**, 145–155.
Winter M, Devictor V, Schweiger O (2013) Phylogenetic diversity and nature conservation: where are we? *Trends in Ecology and Evolution*, **28**, 199–204.
Yang Z (1994) Estimating the pattern of nucleotide substitution. *Journal of Molecular Evolution*, **39**, 105–111.
Yang Z, Rannala B (1997) Bayesian phylogenetic inference using DNA sequences: a Markov chain Monte Carlo method. *Molecular Biology and Evolution*, **14**, 717–724.

H.C. and H.K. conceived the idea of the study and evaluated the results. K.S. designed the impingement fish experiment and collected the data. H.K. formulated the statistical models. H.C. performed the analyses. H.C. and H.K. wrote the manuscript. H.C., K.S. and H.K. checked the manuscript.

Data accessibility

The Supplementary Table lists the species in the impingement meta-sample. The raw data of monthly observations and the Accession nos of the COI sequences are available as supporting data. The sequence data, the estimated tree and the program used for the analysis are attached as a zip file.

Supporting information

Additional supporting information may be found in the online version of this article.

Data S1 The raw data of monthly observations, sequence data and R program.

Fig. S1 Phylogenetic skew index values calculated under different sampling scenarios.

Fig. S2 The family names of the selected species in Fig. S1a.

Fig. S3 Average species composition.

Fig. S4 Average species composition over the years and months from 2001 to 2012.

Table S1 Checklist of species in the meta-community recorded in the impingement samples.