Doctoral Dissertation

博士論文

Development, test and application of a new intra-taxon sampling method based on geographic information

（空間情報に基づく種内新規サンプリング手法の

開発・検証・適用）

A Dissertation Submitted for the Degree of Doctor of Philosophy
December 2019

令和元年 12 月博士（理学）申請

Department of Biological Sciences, Graduate School of Science,
The University of Tokyo
東京大学大学院理学系研究科
生物科学専攻

Satoshi Aoki

青木　聡志

Copyright note

Chapter 2:

Used with permission of John Wiley & Sons publications, from [Where wild species be sampled? New method based on isolation-by-distance objectively gives the answer, Aoki S. & Ito M., Molecular Ecology Resources 20(5), 2020]; permission conveyed through Copyright Clearance Center, Inc.

Chapter 4:

Reproduced with permission from copyright holder. The reviewed and published version of this chapter is the following: Aoki S., Li P., Matsuo A., Suyama Y. & Ito M. (2023) Molecular phylogeny and taxonomy of the genus *Nanocnide* (Urticaceae) with particular attention to the Ryukyu Islands endemic *N. lobata*. Phytotaxa 607(1): 23-40. doi.org/10.11646/phytotaxa.607.1.3

Contents

Abstract

Chapter 2: When estimating nature of wild species, objective sampling is necessary. However, while inter-population or inter-taxa sampling methods have been developed, there are currently few intra-taxon sampling methods to objectively decide where to sample wild taxa. I suggest an alternative to conventional haphazard samplings. The method computes appropriate sampling locations from coordinates, assuming geographical autocorrelation of phylogeny within a species (isolation-by-distance). The computed locations encompass the highest diversity, providing a genetically representative sample. I tested this using published phylogeographical data. The test result was generally encouraging, but the method failed where the species under consideration showed uniform genetic structure or recent distribution expansion, either of which violates the assumption of geographical autocorrelation of phylogeny. Though simple, the new method constructs a methodological and statistical foundation for sampling wild taxa, and is applicable to taxonomy and conservation biology.

Chapter 3: Effect sizes of the difference, or standardized mean differences, are widely used for meta-analysis or power-analysis. However, common effect sizes of the difference such as Cohen's d or Hedges' d assume variance equality that is fragile and is often violated in practical applications. Based on Welch's t tests, I defined a new effect size of the difference between means, which did not assume variance equality, thereby providing a more accurate value for data with unequal variances. In addition, I presented the unbiased estimator of an effect size of the difference between a mean and a known constant. An R package is also provided to compute these effect sizes with their variance and confidence interval.

Chapter 4: *Nanocnide* is composed of three or four herbaceous species, and endemic to East Asia and Vietnam. Apart from recently reported *N. zhejiangensis*, three species are recognized in Japan: *N. japonica*, *N. lobata* and *N. pilosa*. On the other hand, Flora of China synonymizes *N. pilosa* under *N. lobata*. To solve this taxonomic contradiction, I conducted phylogenetic analyses and discussed the taxonomic status of *N. pilosa* and *N. lobata*. The phylogenetic analysis based on nuclear internal transcribed spacer (ITS) and multiplexed ISSR genotyping by sequencing (MIG-seq) showed that *Nanocnide* was divided into three groups which correspond to *N. japonica*, *N. lobata* and *N. pilosa*. While monophyly of *N. japonica* and that of *N. lobata* were confident, the group of *N. pilosa* was paraphyletic (ITS) or a poorly supported monophyly (MIG-seq). Therefore, I suggested to treat *N. pilosa* as a subspecies of *N. lobata* based on cladistic species concept.

Chapter 1. General Introduction

Development of molecular phylogenetics has greatly improved objectivity and reproducibility of taxonomic studies. However, the other conventional methodologies in taxonomy as well are need to be improved for higher objectivity and reproducibility.

Statistics is important in a kind of taxonomic studies. In taxonomy, studies which describe wild species (hereafter referred to as wild taxa) can be roughly divided into two groups: “new descriptions” and “revisions”. The new descriptions report taxa which are newly found. This corresponds to the short research paper in the classification of taxonomic publications by Narendran (2006). New descriptions are often based on limited information on the distribution. The other kind of taxonomic studies, revisions, examine known taxa and change their status if necessary. This corresponds to the revisions and the monographs in Narendran (2006). Revisions usually utilize accumulated information on the distribution. Although some revisions reveal new cryptic taxa, even such new cryptic taxa can be described with plenty of information on their distributions. While there is little room to use inferential statistics in new descriptions, inferential statistics can and should play a great role in revisions.

Statistics is important in revisions, and so is sampling in statistics. However, sampling methods have been paid little respect in revisions; the locations where biological samples are taken have been empirically decided. Since there have been no studies on sampling methods in taxonomy to the best of my knowledge, this convention seems common among taxonomists.

To improve this subjective sampling of wild taxa, I aimed to theorize the conventional sampling method and developed an objective sampling method. This new sampling method is completely different from ordinal random sampling in which all samples are extracted at an equal probability from the statistical mother population, and tries to estimate only the range of the genetic diversity of wild taxa. Such sampling can be realized by maximizing sampled diversity of a wild taxon. In practice, I theorized the diversity-wise sampling of wild taxa and presented a novel sampling method, which objectively decide sampling locations and makes a foundation for future development of statistics based on diversity-wise sampling.

This thesis comprises of three main parts. Chapter 2 introduces the new sampling method for wild taxa. Chapter 3 defines new effect size statistics which are necessary to evaluate the results of the experiment in Chapter 2. Chapter 4 demonstrates a taxonomic study, which can be an example of application of the new sampling method defined in Chapter 2. The concluding remarks follow these main chapters. The appendix

includes supporting figures for Chapters 2 and 4 and the proofs of mathematical theories described in Chapter 3.

Chapter 2. New Sampling Method

2.1 Introduction

When taxonomists try to accurately estimate nature of a wild taxon, objective sampling is necessary, since sampling is a basis of statistical estimation. However, revising taxonomic studies have employed empirical sampling. While many intra-population or inter-taxa sampling methods have been developed, few methods have been invented for intra-taxon sampling. In the field of conservation biology, Quijano et al. (2012) proposed an intra-taxon sampling method which decides sampling locations based on environmental information. However, the basal theory which estimates genetic variation from environmental variation is not formulated (Greene and Hart 1999; Faith and Walker 1996). An alternative sampling method is necessary to improve the conventional haphazard sampling and groundless statistical estimation on wild taxa.

2.2 Materials and Methods

2.2.1 Outline of the new method

Here, I present a new sampling method for wild taxa called “spatial sampling,” which theoretically provides the sampling locations with the highest haplotype diversity from the candidate locations. The candidates of the sampling locations must contain their coordinates. They are supposed to be obtained from known distribution data, but distributions expected by species distribution models may be used. Spatial sampling is based on the idea that “the most efficient sampling” for the limited number of sampling locations is the sampling that achieves the most genetically diverse samples from the candidates. This idea is quite different from and incompatible with that of random sampling. Even though the samples collected using this method cannot estimate ordinary parameters based on the frequency of individuals, the samples can be used to efficiently represent its diversity. The main assumption of spatial sampling is isolation-by-distance (IBD) or the geographical autocorrelation of genetic correlation, meaning that geographically closer individuals should be genetically closer at equilibrium. This has been formulated in population genetics (Malécot 1955; Kimura & Weiss 1964; Weiss & Kimura 1965). Based on this assumption, “the most efficient sampling” is approximately realized as the spatially widest and most uniform combination of locations for the candidates. Moreover, the spatial sampling software has a variable called “necessity” for each candidate. This variable is used to force the software to include some important candidate locations, such as the type locality or environmentally abnormal locations, into the sampling locations. This variable also

enables adaptive sampling (Thompson & Seber 1966), in a broad sense, by optimizing the sampling plan in the course of on-going sampling, depending on the results so far obtained (Fig. 2.1).

2.2.2 Details of the new method

How can one place points widely and uniformly on a sphere? This is an old and unresolved mathematical problem called Tammes's problem, which Tammes proposed in a study on the pores on pollen grains (Tammes 1930). Recently, this problem has come to be studied from the viewpoint of arranging electrons, which repel each other by Coulomb force, on a sphere. Solving Tammes's problem results in finding the position of electrons which minimizes the Coulomb potential energy

$$\mathrm{U}_c(\mathrm{n}) = \sum_{p \neq q}^{n} d_{pq}^{-1},$$

where d_{pq} is the distance between points P and Q on a sphere and n is the number of points (Melnyk et al. 1977; Saff & Kuijlaars 1997). Now, I calculate the genetically most diverse locations on the earth by applying this framework, and for this purpose, I introduce an equation which corresponds to Coulomb force in the above equation. Consider two locations P and Q on a discrete Euclidian two dimensions, a stepping-stone model in Kimura & Weiss (1964). These locations are separated by k_1 steps in the X axis and k_2 steps in the Y axis, and both locations have biological populations of a species with the same population sizes. Then, I consider an allele A and its frequencies in the populations at P and Q are denoted as p_P and p_Q, respectively. Likewise, its frequency among all the locations in the two-dimensional space is denoted as $\bar{p}$. Individuals at a given location have a chance of migration to the four adjacent locations at the rate of $m_{1/2}$ per generation. They also have a chance of long-distance migration to all the locations except for the original one at the rate of m_∞. Now, when $\rho = \sqrt{{k_1}^2 + {k_2}^2}$, the gene correlation $r(\rho)$ between populations at the locations P and Q is

$$r(\rho) = \frac{E[(p_P - \bar{p})(p_Q - \bar{p})]}{E[(p_P - \bar{p})^2]}.$$

The gene correlation increases when p_P and p_Q get similar. In other words, higher gene correlation between two locations means lower diversity among them. Here, the gene correlation is known to decrease when the distance between these two locations increases (Kimura & Weiss 1964; Weiss & Kimura 1965). Assuming $m_1 \gg m_\infty$, the gene correlation $r(\rho)$ between two locations separated by distance ρ on a discrete

two-dimensional plane at equilibrium is proportional to

$$e^{-a\rho}/\sqrt{\rho},$$

where

$$a = \sqrt{4m_\infty/m_1}.$$

Kimura & Weiss (1964) also considered situations in one- and three-dimensional spaces, and Malécot (1955) considered continuous multi-dimensional space. What might be a problem in application of this theory is that none of them considered a sphere surface which is looped, non-Euclidian and similar to the surface of the Earth. However, all the above studies (Kimura and Weiss1964; Malécot 1955) obtained similar results that the gene correlation between two locations approximately exponentially decreases as the distance between the two locations increases. Therefore, although there might be minor inaccuracies in the coefficients, the above relationship in two-dimension by Kimura & Weiss (1964) was employed for application, because their work (Kimura & Weiss 1964) was described more minutely and insisted to be more accurate than Malécot (1955) by them. When applying the above relationship to finding locations with the most diverse populations of a wild taxon, I assume a is equal to or smaller than 1 because $m_1 \gg m_\infty$ is assumed in the above relationship. For this reason and the potential inaccuracy in the coefficient when applying to a sphere surface, $a = 1$ is tentatively assumed in this practical application. The gene correlation represents only the diversity between two locations. To quantify the diversity among n locations, I utilize the sum of the gene correlations $\mathrm{U(n)}$ for the whole locations:

$$\mathrm{U(n)} = \sum_{p \neq q}^{n} r(d_{pq})$$

$$= \sum_{p \neq q}^{n} e^{-d_{pq}}/\sqrt{d_{pq}}.$$

Therefore, the combination of locations that minimize this $\mathrm{U(n)}$ represents the samples with the highest haplotype diversity. For the sum of the gene correlation, the distance between locations should not be a Euclidean distance but a spherical one, because wild taxa are practically thought to move on the surface of the earth.

Spatial sampling assumes geographical autocorrelation of gene correlation (IBD), genetic equilibrium, uniform mutation rate, and uniform dispersal ability. On the other hand, spatial sampling does not assume environmental information including geographical barriers.

2.2.3 Software and algorithm for the new method

I present Samploc, the software to conduct spatial sampling with a graphical user interface. Samploc is written in C# and available for Windows 7 or later. The usage of Samploc is summarized in Fig. 2.1. Samploc can import coordinates expressed in DEG (e.g. 35.51), DMS (N35°30'36"), and DMM (N35°30.600'), and can handle these different expressions co-used in a single file. Also, Samploc can adapt to irregular units of coordinates, such as “o” used as the unit of degree, by setting the option. Thus, Samploc can easily import coordinates data created by other researchers or institutes.

Samploc assumes the earth to be a sphere or an ellipsoid of revolution (a warped sphere). The distance on an ellipsoid of revolution is calculated with GeographicLib (Karney 2013), and almost any Earth ellipsoid can be used by inputting the parameters.

The gene correlation U(n) is expressed as the sum of the powers of the base of the natural logarithm (Napier’s constant e). If a distance expressed in kilometers is used for the calculation, the gene correlation for the two antipodal points on the earth cannot be calculated. This is because the resultant number becomes too small to treat as an ordinal floating-point data type (double) in a computer. Therefore, the gene correlation is calculated based on the distance expressed in 100-kilometer units for the earth. This distance unit is automatically decided based on the parameters of the ellipsoid. In the calculation of the gene correlation, the Kahan summation algorithm (Kahan 1965) was employed to reduce the summation error.

Even though solving Tammes’s problem for all the possible combinations of locations is difficult, it is relatively easy to solve it for a finite number of candidate locations. Samploc requires inputting a number which indicates how many combinations of possible sampling locations Samploc examines during calculation, and I refer to this number as a searching number. Samploc selects its calculation methods from two ways based on these numbers. When the number of possible combinations is not larger than the searching number, Samploc conducts an exhaustive search; it checks all the combinations and returns the best one. By the way, the number of possible combinations grows drastically as the number of candidates increase. This combinatorial explosion makes the exhaustive search time-consuming or even impossible to conduct in an ordinal computer. When the number of possible combinations is larger than the searching number, or when the number of possible combinations is more than the maximum integer in the ordinal integer data type (int), Samploc conducts a heuristic search. Precisely, it conducts the simulated annealing (Khachaturyan et al. 1979) and provides a good combination even for numerous possible combinations.

The outline of the simulated annealing is the following: The simulated annealing employed in Samploc is a repeating procedure, and it always has a "temperature" parameter, a combination of locations and the score of the combination. The score here means the sum of the gene correlations. First, set a "temperature" parameter and obtain a random combination of locations and its score. By exchanging a location in the combination for a location which is not in the combination, a new combination and its score are obtained. Then, comparing the current score and the new score, a system of simulated annealing accepts the exchange of the locations at a particular probability. This probability is decided by the "temperature" and difference between the new score and the current score; the probability is higher when the "temperature" is high and when the difference of scores is large. If the new combination is accepted, the next new combination of locations is generated from the new combinations. Otherwise the new combination and its score are abandoned, and the next new combination of locations is generated from the current combination. The "temperature" always decreases irrespective of the result of the acceptance. This procedure means that the status of simulated annealing always tends to move to combinations with better score, but when the procedure is young and the "temperature" is high, the status tends to permit to move to combinations with worse scores. This enables the status moving to the optimal solution and evacuating from local optimal solutions.

Details of the simulated annealing algorithm in Samploc are the following: The simulated annealing is based on Xorshift pseudorandom numbers which has an advantage in quick calculation (Marsaglia 2003). The starting "temperature", T_0 is 10, and the "temperature", T_{r+1} after r+1 repeats of the simulated annealing is

$$T_{r+1} = T_r(0.001/T_r)^r.$$

The probability P ($0 \leqq P \leqq 1$) at which the exchange of locations is accepted in *r*th repeat is

$$P = 1 - currentScore * T_{r-1}/candidateScore,$$

where "currentScore" is the sum of gene correlation described above before the exchange and "candidateScore" is that after the exchange. When exchanging locations, any location with its necessity variable, "necessary" is always kept in the system's combinations and is not exchanged. Locations whose necessity variable is "unnecessary" are not included in the system's combination, and only the locations whose necessity variable is "normal" (default) can be the candidate. The exchanges are conducted in two ways, and these two exchanging methods are used alternately. One is the random

exchange. This literally means exchange a random location in the system for a randomly selected one out of the system. Another method is to replace one of the two geographically nearest locations with a randomly selected location out of the system, because the nearest pair gives the worst influence on the current score. The location to be replaced is selected at random from the nearest two locations. The second exchanging method has much more chance to improve the current score than the first random exchanging method, but, because of this nature, it is more likely to be trapped in local optimal solutions. Therefore, Samploc employed both of them as a compromise. These parameters were adjusted using existing phylogeographical data described in the next section.

2.2.4 Test of the new method

The assumptions of spatial sampling are simple; consequently, it ignores many other factors, such as the effect of geographical barriers. Therefore, it is desirable to test spatial sampling against wild taxa. However, such a test is actually impossible. This is because this requires a mother population of wild taxa from which samples are resampled in spatial sampling and the other counter method, but no one can sample all the individuals of a wild taxon to obtain the mother population. Therefore, I conducted an alternative test using published phylogeographical data for 20 taxa as the mother populations (Table 2.1). I used data from studies which sampled 10 or more locations and from which I could reproduce the coordinates paired with the sequence data. I resampled 20, 40, 60, 80 and 100% of the sampling locations using spatial sampling and random sampling in which all the locations are selected at an equal probability (e.g., Fig. 2.2). In this test, the distance between locations was calculated based on the World Geodetic System 1984 ellipsoid, which is used in Global Positioning Systems (GPS). Random resampling was conducted 20 times for each resampling percentage using Mersenne twister pseudorandom numbers (Matsumoto & Nishimura 1998), which is slower than Xorshift, but generates pseudorandom numbers with better quality. For spatial sampling, the searching number was set to 10,000 and repeated 20 times for each resampling percentage, except when the total number of possible combinations was 200,000 (= 20×10,000) or smaller. For example, the all combinations for resampling 10 locations out of 50 candidates is ${}_{50}C_{10}$ = 10,272,278,170 > 200,000, and a simulated annealing was employed in such cases. When the combination was 200,000 or smaller, the best solution was calculated only once by exhaustive search. While the genetic diversity among locations is measured in terms of the sum of the gene correlations, this measure is not standardized in a specific range like 0–1, resulting in difficulty in

comparison of data from different numbers of locations. Ideally, the comparing diversity should be non-frequentistic one, but it is undefined and its definition requires diversity-wise sampling which this study treats. Therefore, for the comparison, I used haplotype diversity (Nei and Tajima 1981), which is popular and standardized, although the method to convert gene correlation or its sum to haplotype diversity is not known. The unbiased haplotype diversity h is defined as

$$h = \{n/(n-1)\}(1 - \sum_i {x_i}^2),$$

where n is the total number of sequences in the population and x_i is the frequency of the ith haplotype (Nei and Tajima 1981). Note that two sequences with one or more substitution(s) are considered as distinct haplotypes. In addition to the haplotype diversity, I also compared the nucleotide diversity, because nucleotide diversity reflects phylogenies which haplotype diversity ignores, and because spatial sampling ideally pursues phylogenetic representativity. The unbiased nucleotide diversity π was calculated based on the original definition (Nei & Li 1979; Nei & Tajima 1981) as,

$$\pi = \{n/(n-1)\} \sum_{ij} x_i x_j \pi_{ij},$$

where n is the total number of sequences, x_i and x_j are ith and jth sequence frequency and π_{ij} is the number of nucleotide differences per nucleotide site between the ith and jth sequences. The other two definitions neither of which was discussed in the original literature are provided here. First, after aligning the sequence with the online version of MAFFT (Katoh & Standley 2013), sites with a gap (-) were excluded from the calculation. Second, mixed bases (including "N") were treated as having equal probability of being the bases which compose the mixed base. In such cases, the distance was calculated as the expected value. For example, R is a mixed base composed of A and G, and was treated as being A at 50% and G at 50%.

I compared the diversities using the rate of increase and the effect size of the difference, because hypothesis tests in simulation studies are not appropriate (White et al. 2013). The used effect sizes of the difference were e and c (see Chapter 3 in this thesis). The effect size e is defined between two means without assuming variance equality, and was employed for the data of the simulated annealing, because the data sometimes had zero variance, which caused the expected variance inequality. The effect size c is defined between a mean and a constant, and was employed for the data of the exhaustive search. I judged results with an absolute effect size smaller than 0.2 as a "non-considerable difference" based on Cohen's standard (Cohen 1988).

I also estimated IBD in the original data to discuss the result. The pairwise

geographic distance, pairwise Gst (Pons and Petit 1995), and pairwise Nst (Pons and Petit 1996) of the original data of each taxon were calculated by the software other than Samploc. Their definitions are the following:

$$h_S = \left\{ \sum_k \left(1 - \sum_i {p_{ki}}^2 \right) \right\} / n$$

$$h_T = 1 - \sum_i {p_i}^2$$

$$G_{ST} = (h_T - h_S)/h_T$$

$$V_S = (\sum_k \sum_{ij} \pi_{ij} p_{ki} p_{kj})/n$$

$$V_T = \sum_{ij} \pi_{ij} p_{ki} p_{kj}$$

$$N_{ST} = (V_T - V_S)/V_T.$$

In the above definitions, p_i is the ith allele or sequence frequency in the total population, p_{ki} is that in the kth subpopulation, n is the number of subpopulations, and π_{ij} is the same as that used for nucleotide diversity. The $\sum_{ij} X$ means the sum of X about all the combinations of sequences (including $i = j$ which just gives 0). However, G_{ST} and N_{ST} in this study were treated as 0, when $h_T = 0$ and $V_T = 0$, respectively, although this treatment was not defined in the original definition (Pons and Petit 1995; Pons and Petit 1996). The Gst and Nst correspond to the haplotype diversity and the nucleotide diversity, respectively. Neither the Gst nor Nst is the unbiased statistics, because many locations in the data had only one individual, which prevented the calculation of the unbiased statistics. Using these data, I conducted simple Mantel test (Mantel 1967) for distance vs. Gst and distance vs. Nst using ade4 R package (Chessel et al. 2004) with the significance level = 0.05. This reveals significance of geographical autocorrelation of genetic distances. Moran's test at subpopulation level (Barbujani 1987) can also be applied to the data for the same purpose. However, it can be applied only to rare alleles seen in a single location (e.g. Iwasaki et al. 2012), and therefore I did not conduct it. In addition, considering advices by Bohonak (2002), I calculated standard(ized) major axis (SMA) regression (also called reduced major axis regression) line for distance vs. Gst and distance vs. Nst using smatr R package (Warton et al. 2011). I did not intend to check the linear relationship among them by this regression analysis, but this analysis was conducted to report whether the correlation was positive or negative by the slope. The test of the slope = 0 cannot be conducted for SMA (Sokal and Rohlf 1995). While Rousset (1997) showed Fst / (1 - Fst) has a linear

relationship with geographical distance in some situations, I did not use corresponding Gst / (1 - Gst) or Nst / (1 - Nst). This is because many of the data had Gst = 1 or Nst = 1, and the assuming situation depends on the habitat (subpopulation) size which is hard to estimate.

The original sampling locations were not uniform, and the number of collected individuals in a location was not fixed. Therefore, this test had a somewhat stochastic nature. However, if spatial sampling gives more genetically diverse samples, as expected, the haplotype diversity from spatial sampling will exceed that from random sampling, and likely so will the nucleotide diversity. The random resampling, calculation of haplotype diversity, nucleotide diversity, and effect sizes were conducted with original Visual Basic for Applications macros. The calculation of pairwise distance, Gst, and Nst was conducted by software I wrote in C#.

2.3 Results

The haplotype diversity increased in 42 pairs, decreased in 27 pairs, and remained unchanged in 16 pairs by spatial sampling compared to random sampling among the 80 comparisons with the exception of the 100% resampling pairs which necessarily produce the same values for both resampling methods (Figs. S1–S20). The nucleotide diversity increased in 55 pairs, decreased in 19 pairs, and remained unchanged in 6 pairs (Figs. S1–S20). Table 2.2 shows the average rates of increase and the effect sizes for each taxon. Judging from the average effect sizes, the haplotype diversity decreased in a monkey, a fish, a mushroom, a bird, two tardigrades, and a brown alga (Figs. S1a, S5a, S7a, S9a, S12a, S13a, and S15a), and remained unchanged in a tree, a sea urchin, and a red alga (Figs. S2a, S14a, and S16a). The nucleotide diversity was decreased in a tardigrade, a brown alga, and a fern (Figs. S13b, S15b, and S18b), and remained unchanged in a mosquito and a tardigrade (Figs. S10b and S12b). Figures S1–S20 show the results for each taxon. The result of tests on IBD was summarized in Table 2.3. IBD was significant in 7 Gst data and 9 Nst data, and spatial sampling was effective in 4 haplotype diversity data and 7 nucleotide diversity data among them. All of these data with significant IBD had positive slopes of the regression lines. On the other hand, IBD was not significant in 13 Gst data and 11 Nst data, and spatial sampling was effective in 6 haplotype diversity data and 8 nucleotide diversity data among them.

2.4 Discussion

2.4.1 On overall result of the test

Here, I discuss general tendency of the result. The discussions on each of these taxa are

provided in the next subsection. Spatial sampling worked well in more than half the comparisons, but not in all of them. Both of the individual comparisons (Figs. S1–S20) and the averaged results (Table 2.2) implied that spatial sampling worked better for nucleotide diversity than haplotype diversity. This may be because when the distance between two locations increases, haplotype diversity in the two locations stops increasing at the time when the locations lose the last common haplotype. In other words, nucleotide diversity is better at reflecting genetic distance between two remote locations than haplotype diversity. All the data with significant IBD had positive line of slopes (Table 2.3). This supports the principle of IBD relationship which spatial sampling relies on. Even when IBD was significant, spatial sampling did not always provide higher diversities, but it seems that spatial sampling for IBD-significant taxa worked better than that for taxa in which IBD was not significant. I dare to note that not-significant IBD does not always mean no IBD.

Possible reasons for the malfunction of spatial sampling are genetic non-equilibrium after the last glacial period (LGP) (observed in a monkey, a tree, a fish, a mosquito, and a red alga), unclear genetic structures (a mushroom, a bird and a brown alga), incorporation of multiple taxa (a mushroom and a fern), recent long-distance dispersal (a fish and a tardigrade), and ignorance of geographic barriers (a sea urchin and a fern). Malfunctions in a tardigrade seemed to be caused by a shortage of sampling locations which did not cover the whole distribution of the taxon. The recent distribution expansion after the LGP violates the assumption that the gene frequency is at equilibrium. Although there are some theories on population expansion (e.g., Nichols & Hewitt 1994; Ibrahim et al. 1996) and models of paleoclimate (e.g., Kageyama et al. 2018), integrating them into spatial sampling will be a complicated task. Geographical barriers can also be considered in the same framework as the LGP problem. However, the impact of ignoring geographical barriers on spatial sampling seems smaller than one might expect. Recent barriers like straits formed after the LGP for land organisms lack enough time to affect the result of spatial sampling just like seen in the result of the monkey or the tree. On the other hand, ancient and strong barriers like mountain range for low land organisms causes speciation, and therefore it does not affect the validity of spatial sampling within a taxon. Only ancient and moderate barriers like Italian peninsula for sea urchins seem to form irregular genetic structure within a taxon without specification, which is not assumed by spatial sampling, although what “moderate” means depends on the researcher’s definition of a taxon. The absence of clear genetic structures means that the dispersal ability of the taxon is high relative to its distribution, and that the whole distribution is approximately assumed to be a single breeding

population. However, no other sampling methods currently cope with it better than the spatial sampling, and I suggest employing spatial sampling even when this situation is expected. Incorporation of multiple taxa is also problematic. However, this seems inevitable when the phylogenetic identity of a taxon is not available in advance. In this case, any sampling must rely on a morphologically identified taxon as the working hypothesis. Recent long-distance dispersals cause unignorable effects on the equilibrium assumption, regardless of whether they are natural or artificial. However, such rare dispersals are difficult to predict. Therefore, they should not be assumed unless already known. When known, such locations can be omitted from or necessarily included into the sampling using the necessity variable in Samploc.

2.4.2 On each taxon in the test

Here, I discuss the taxa which did not show increased haplotype or nucleotide diversity in spatial sampling. *Macaca fuscata*, a monkey, showed decreased haplotype diversity, whereas its nucleotide diversity increased (Fig. S1). While IBD was significant for Gst, the northern individuals had only a single haplogroup (Fig. S1; Kawamoto et al. 2007). This was caused by a recent distribution expansion after the LGP, and may have triggered fewer haplotypes in spatial sampling and, therefore, lower haplotype diversity. *Fagus crenata*, a beech wood, showed an unchanged haplotype diversity (Fig. S2a), and this is probably due to similar reasons to the monkey. *Oryzias latipes*, a ricefish, showed decreased haplotype diversity (Fig. S5a). This may be due to a recent distribution expansion after LGP in Seto Inland Sea and gene flows from western Japan to Kanto region by human activity (Takehana et al. 2003). *Amanita muscaria* Clade I, a mushroom, also showed decreased haplotype diversity (Fig. S7a). I conducted phylogenetic analysis on this clade, because the phylogenetic tree in the original study (Geml et al. 2008) was based on multiple regions which were not extracted from all the samples and not used in spatial sampling. As a result, this clade could be divided into three subclades composed of Alaskan samples, Mexican and Arizonan samples, and the other samples from North America in the middle latitude (Fig. S21). The subclade in the middle latitude did not show a clear genetic structure, probably because of its high dispersal ability relative to the wideness of its distribution. This violates the assumption of geographical autocorrelation of phylogeny. Despite this dispersal ability, these three subclades had clearly separate distributions. Therefore, they might correspond to independent biological (breeding) species. In short, the decrease seemed to be caused by violation of the assumption of autocorrelation and incorporation of multiple species in the analysis. *Ptionorhynchus violaceus violaceus*, a bird, showed lower haplotype diversity (Fig. S9a). The 20% resampling showed an especially large decrease in both

haplotype diversity and nucleotide diversity (Fig. S9b). This taxon also does not have a clear genetic structure, which was supported by the highest p-values of all taxa in Mantel test of Gst and Nst (Table 2.3). Therefore, the assumption of geographical autocorrelation of phylogeny is not fulfilled for this taxon. *Wyeomyia smithii*, a mosquito, showed unchanged nucleotide diversity (Fig. S10b). This was caused by a decrease in the nucleotide diversity in the 60% resampling which stemmed from the northern individuals being genetically close relative to the distance, due to recent expansion after the LGP. *Echiniscoides sigismundi*, a tardigrade, did not show increases in either haplotype diversity or nucleotide diversity (Figs. S12 and S13). Although the IBD was significant for Nst, the northern taxon has some locations with a rare combination of haplotypes. The authors of the original article pointed that these locations were supposedly formed by rare trans-Atlantic dispersal or dispersal by ballast water (Faurby et al. 2011), and spatial sampling did not include these points because of their position. More original sampling locations might improve the results of spatial sampling. In the southern taxon, one location had a unique and distantly related haplotype, even though it was not geographically distant from the other locations. The authors guessed that these haplotypes originated at a southern refugium during the LGP (Faurby et al. 2011). If the sampling included more southern regions, spatial sampling might work well for this taxon. *Paracentrotus lividus*, a sea urchin, showed unchanged haplotype diversity in spite of the significant IBD in Gst. This was mainly caused by a low haplotype diversity in the 20% resampling (Fig. S14). In the 20% resampling by spatial sampling, the single optimal combination of locations was calculated, and it lacked samples from the Adriatic Sea, which was relatively rich in a unique haplogroup (Maltagliati 2010). In short, the low haplotype diversity was caused by the method's ignorance of the geographical barrier which, in this case, corresponded to the Italian Peninsula. *Sargassum horneri*, a red alga, also showed a decrease in both diversity indexes (Fig. S15). This is because one haplogroup (Clade 2-1 in Uwai et al. 2009) became widely distributed during expansion after the LGP. *Gelidium lingulatum*, a brown alga, had unchanged haplotype diversity (Fig. S16). The genetic structure of this taxon did not have a geographical pattern due to its long-distance dispersal (López et al. 2017). Finally, the *Pteridium aquilinum*, a fern, showed decreased nucleotide diversity despite the significant IBD for Nst (Fig. S18b). This taxon has several subclades within it, such as ones in Africa-Europe, South to Southeast Asia, North to East Asia, or North America. Spatial sampling seems to fail balanced sampling of these subclades which are not necessarily distributed in a geographically uniform manner. Therefore, this problem can be expressed in two ways: Incorporation of multiple taxa corresponding to the

subclades, or ignorance of geographical barriers such as the Himalayan range, Pacific Ocean, or Atlantic Ocean.

2.4.3 On points other than the test

I have some other points to mention. First, spatial sampling requires candidate locations, and the known distribution of the wild taxon is used for the candidates. When one has no information on the distribution, potential sampling locations can be the candidates. Second, spatial sampling does not decide the number of samples in a location, and it depends on the type of study. Third, walking distance within a sampling location is usually short and negligible comparing to the distances between the calculated sampling locations.

This study replaces the current haphazard sampling of wild taxa which cannot be a basis of statistical estimation. Generally, random sampling is used for statistical estimation, although it is difficult to apply to wild taxa (Zar 2014; Sokal & Rohlf 2012). Random sampling and spatial sampling are totally different. While random sampling extracts samples at an equal probability from the mother population, spatial sampling extracts samples which are estimated to have the highest genetic diversity.

Even though the new method is still simple and representative statistics for wild taxa based on spatial sampling are lacking, I hope this spatial sampling will promote the development of a statistical foundation for field biologists sampling wild taxa.

2.5 Figures and Tables

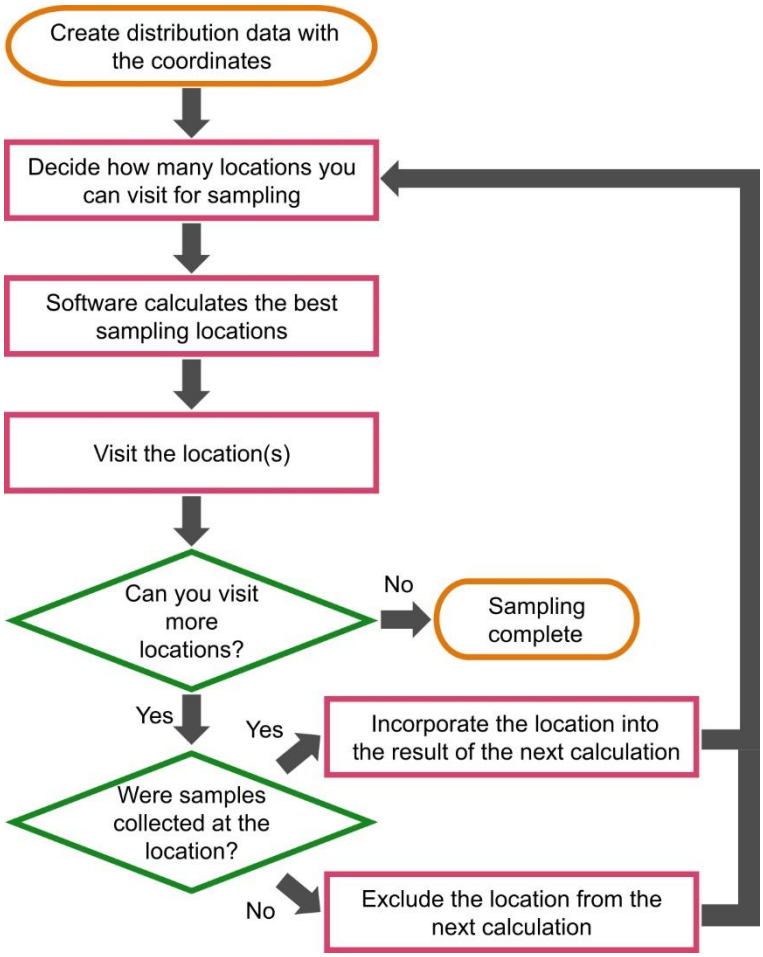

Figure 2.1. Flowchart of spatial sampling. First, create distribution data with the coordinates of the wild taxon as a comma-separated values (csv) file. Each location in the file has a necessity variable which is usually "normal". However, if one believes some locations are so important that they must be sampled (such as the type locality), change the necessity variables to "necessary." Then, decide how many locations one can visit, considering time and budget constraints. Next, by inputting this number and the csv file, the software calculates the best combination of sampling locations. Now, visit some of them. When one cannot visit any more locations, the sampling ends. Otherwise, check whether one could collect the samples at the visited location(s). If one could, change the necessity variable of the location to "necessary." This enables one to include the location into the result of the next calculation. If one could not collect the samples at the visited location(s), the necessity variable of the location must be changed to "omitted" to exclude the location from the next calculation. After that, reconsider the number of locations one can visit, and continue this loop until one's time or budget runs out.

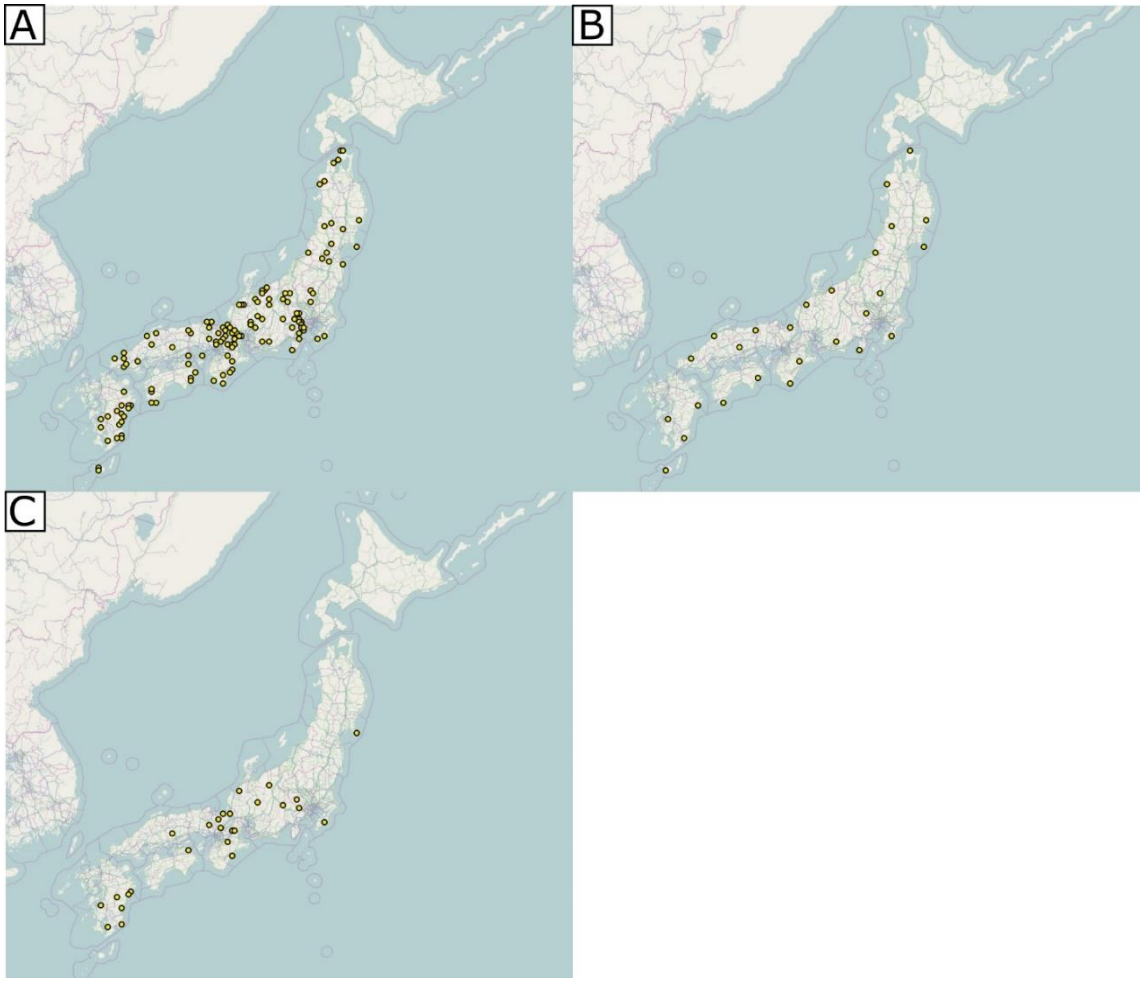

Figure 2.2. Candidates for sampling location and the results of the two sampling methods. (A) The 131 sampling locations of the *Macaca fuscata* monkey used in the published study (Kawamoto et al. 2007). (B) The 26 locations resampled from the candidates in (A) using spatial sampling. (C) The 26 locations resampled from the candidates in (A) using random sampling. Map data: ©OpenStreetMap contributors, license: http://opendatacommons.org/licenses/dbcl/1.0/

Table 2.1. Phylogeographical studies used for the experiment. The "n" in the first row stands for the number of sampling locations. Bracketed letters in "Used Region(s)" stand for the source organelles of the sequences.

Taxon	Group	Reference	n	Used region(s)
Macaca fuscata	Monkey	(Kawamoto et al. 2007)	131	mitochondria control region (mt)
Fagus crenata	Tree	(Fujii et al. 2002)	45	trnL-trnF, trnK (cp)
Mauremys japonica	Turtle	(Suzuki & Hikida 2011)	43	cytochrome b (mt)
Oryzias sakizumii (Clade A in reference)	Fish	(Takehana et al. 2003)	18	cytochrome b (mt)
Oryzias latipes (Clade B+C in reference)	Fish	(Takehana et al. 2003)	56	cytochrome b (mt)
Parides alconius alconius	Butterfly	(Kato & Yagi 2004)	23	NADH[1] dehydrogenase subunit 5 (mt)
Amanita muscaria Clade 1 in the reference	Mushroom	(Geml et al. 2008)	21	internal transcribed spacer (nc)
Amanita muscaria Clade 2 in the reference	Mushroom	(Geml et al. 2008)	31	internal transcribed spacer (nc)
Ptionorhynchus violaceus violaceus	Bird	(Nicholls & Austin 2005)	23	ATP[2]ase 6, ATPase 8 (mt)
Wyeomyia smithii	Mosquito	(Emerson et al. 2010)	15	COI[3] (mt)
Francisella tularensis subsp. *tularensis*	Bacterium	(Vogler et al. 2009)	23	canonical SNP[4] (cy)
Echiniscoides sigismundi SigiNorth	Tardigrade	(Faurby et al. 2011)	17	COI (mt)
Echiniscoides sigismundi SigiSouth	Tardigrade	(Faurby et al. 2011)	10	COI (mt)
Paracentrotus lividus	Sea urchin	(Maltagliati et al. 2010)	26	cytochrome b (mt)

Sargassum horneri	Brown alga	(Uwai et al. 2009)	24	COIII[5] (mt)
Gelidium lingulatum	Red alga	(López et al. 2017)	20	COI (mt)
Gelidium rex	Red alga	(López et al. 2017)	11	COI (mt)
Pteridium aquilinum	Fern	(Der et al. 2009)	61	trnS-rpS4, rpL16 intron (cp)
Pteridium aquilinum subsp. *aquilinum*	Fern	(Der et al. 2009)	17	trnS-rpS4, rpL16 intron (cp)
Metaphire sieboldi	Earthworm	(Minamiya et al. 2009)	60	COI & 16S ribosomal DNA (mt)

(cp) = chloroplast. (mt) = mitochondrion. (nc) = nucleus. (cy) = cytosol.

[1]NADH: the reduced form of nicotinamide adenine dinucleotide

[2]ATP: adenosine triphosphate

[3]COI: cytochrome oxidase subunit I

[4]SNP: single nucleotide polymorphism

[5]COIII: cytochrome oxidase subunit III

Table 2.2. Average effect size of the difference in diversity for each taxon using spatial sampling relative to random sampling. The 100% resampled data were excluded from these averages.

Taxon	Group	Haplotype diversity Effect size	Nucleotide diversity Effect size
Macaca fuscata	Monkey	-2.2	0.67
Fagus crenata	Tree	-0.10	1.6
Mauremys japonica	Turtle	0.61	0.60
Oryzias sakaizumii	Fish	0.97	1.2
Oryzias latipes	Fish	-0.20	1.6
Parides alconius alconius	Butterfly	0.80	0.57
Amanita muscaria Clade 1	Mushroom	-0.73	1.2
Amanita muscaria Clade 2	Mushroom	2.2	1.9
Ptilonorhynchus violaceus violaceus	Bird	-1.2	0.41
Wyeomyia smithii	Mosquito	1.0	0.13
Francisella tularensis subsp. *tularensis*	Bacterium	0.75	1.2
Echiniscoides sigismundi North	Tardigrade	-0.41	-0.10
Echiniscoides sigismundi South	Tardigrade	-0.59	-0.75
Paracentrotus lividus	Sea urchin	0.19	0.88
Sargassum horneri	Brown alga	-3.4	-2.9
Gelidium lingulatum	Red alga	-0.18	0.47
Gelidium rex	Red alga	0.75	1.1
Pteridium aquilinum	Fern	1.2	-1.5
Pteridium aquilinum subsp. *aquilinum*	Fern	0.35	0.29
Metaphire sieboldi	Earthworm	1.5	2.5

Table 2.3. Test results of isolation by distance.

		Mantel test p-value		Regression slope	Regression slope
Taxon	**Group**	**Gst**	**Nst**	**Gst**	**Nst**
Macaca fuscata	Monkey	1.0E-04	1.0E-04	6.5E-04	6.3E-04
Fagus crenata	Tree	1.0E-04	1.0E-04	9.9E-04	1.0E-03
Mauremys japonica	Turtle	8.4E-03	1.0E-04	1.2E-03	1.4E-03
Oryzias sakaizumii	Fish	2.8E-01	7.2E-01	-1.1E-03	7.1E-04
Oryzias latipes	Fish	1.2E-01	8.9E-02	4.7E-04	5.0E-04
Parides alconius alconius	Butterfly	3.0E-01	3.0E-01	1.5E-03	1.5E-03
Amanita muscaria Clade 1	Mushroom	3.1E-01	3.2E-01	-1.7E-04	-1.6E-04
Amanita muscaria Clade 2	Mushroom	9.0E-04	9.0E-04	1.2E-04	1.2E-04
Ptilonorhynchus violaceus violaceus	Bird	1.0E+00	9.8E-01	-6.5E-04	-6.5E-04
Wyeomyia smithii	Mosquito	9.7E-02	9.3E-02	3.4E-04	2.7E-04
Francisella tularensis subsp. *tularensis*	Bacterium	1.3E-01	9.9E-02	2.7E-04	2.7E-04
Echiniscoides sigismundi North	Tardigrade	4.8E-01	1.0E-04	-5.7E-05	1.1E-04
Echiniscoides sigismundi South	Tardigrade	7.6E-01	8.9E-01	-7.7E-05	-7.4E-05
Paracentrotus lividus	Sea urchin	2.9E-03	8.3E-03	1.6E-05	9.8E-05
Sargassum horneri	Brown alga	4.7E-01	4.6E-01	-7.8E-04	-8.0E-04
Gelidium lingulatum	Red alga	2.2E-01	1.3E-01	-6.8E-04	7.3E-04
Gelidium rex	Red alga	6.9E-03	4.5E-03	2.2E-03	2.3E-03
Pteridium aquilinum	Fern	1.0E-04	1.0E-04	6.8E-05	6.8E-05
Pteridium aquilinum subsp. *aquilinum*	Fern	9.2E-01	9.3E-01	-4.7E-04	-4.7E-04
Metaphire sieboldi	Earthworm	9.9E-01	9.5E-01	1.4E-03	1.3E-03

Chapter 3. New Effect Sizes

3.1 Introduction

Effect sizes of the difference or, more precisely, standardized mean differences between two groups, are widely used to estimate the magnitude of effect independent of the sample size (Nakagawa & Cuthill 2007), to conduct meta-analysis (Glass 1976), or to conduct power-analysis (Cohen 1969). The American Educational Research Association (AERA) or the American Psychological Association (APA) strongly recommend effect sizes are reported in the corresponding fields (AERA 2006; APA 2009). Furthermore, the misuse and misunderstanding of p-value have become public (Wasserstein & Lazar 2016), and use of effect sizes is spreading beyond pedagogy and psychology, where effect sizes have developed, into areas such as biology (Nakagawa & Cuthill 2007). In spite of such importance, the classical effect sizes of the difference assume variance equality (homoscedasticity), which is hard to assume practically or is even expected to be violated a priori in clinical data (Grissom 2000). While Bonett (2008) defined a confidence interval of an effect size estimator which did not assume homoscedasticity, its parameter was not defined. This problem of variance inequality (heteroscedasticity) has been long debated (Grissom 2001; Marfo & Okyere 2019). In addition, the unbiased estimator of an effect size of the difference between a mean and a constant was undefined. To solve these problems, based on Welch's *t* test (Welch1938; Welch 1947), we defined an effect size of the difference between means that does not assume homoscedasticity and calculated the unbiased estimator of an effect size of the difference between a mean and a constant. Effect size of the difference was developed by Cohen (1962), who studied in the field of psychology. Cohen (1962, 1969) defined the effect size as a parameter for two independently and normally distributed populations, $Y^1 \sim N(\mu_1,\ \sigma^2)$ and $Y^2 \sim N(\mu_1,\ \sigma^2)$:

$$\delta = (\mu_1 - \mu_2)/\sigma, \quad (1)$$

which is expressed as d in the original articles (Cohen 1962; Cohen 1969). Note that both populations share the common variance σ^2. The estimator of this parameter was represented as d_s in (Cohen 1969). However, we refer to this estimating statistic as g to distinguish it from the other d we introduce later. The statistic g is defined as

$$g = (\bar{Y}^1 - \bar{Y}^2)/S^{pooled}, \quad (2)$$

where

$$S^{pooled} = \sqrt{\frac{s_1^2(n_1-1)+s_2^2(n_2-1)}{n_1+n_2-2}},$$

and, for $i = 1, 2,$

$$s_i^2 = \frac{\sum_{j=1}^{n_i}(Y_j^i - \bar{Y}^i)^2}{n_i - 1}. \text{ (3)}$$

Here, $\bar{Y}^1$, Y_j^1 and n_1 are the mean of the sample, the sample (random variable), and the sample size of group 1, respectively, while $\bar{Y}^2$, Y_j^2 and n_2 are those of group 2. For the denominator, this effect size uses the pooled standard deviation, which suggests the most precise population variance under the assumption of equal variance (Hedges 1981).

In the field of pedagogy, Glass (1976) suggested another effect size of the difference, independently of Cohen's works. He defined it as "the mean difference on the outcome variable between treated and untreated subjects divided by the within group standard deviation," where "the within groups standard deviation" corresponds to the standard deviation of the untreated group. He clearly distinguished the treated (experimental) group from the untreated (control) group, and there was no assumption regarding the two groups. His effect size was subsequently formulated and named Glass' Δ by Hedges (1981), which is

$$\Delta = (\bar{Y}^E - \bar{Y}^C)/S^C, \text{ (4)}$$

where $\bar{Y}^E$ is the mean of the variable in the experimental group, $\bar{Y}^C$ is that in the control group, and S^C is the unbiased standard deviation of the control group. Hedges (1981) also defined the δ (1) and the g (2) independently of Cohen. Furthermore, Hedges (1981) indicated that g (2) is biased from δ (1), making it unsuitable for analyses that do not treat the entire population. The unbiased estimator of δ (1) is defined as g^U in (Hedges 1981) and d in Hedges & Olkin (1985). In this study, we call it d, which is

$$d = J(n_1 + n_2 - 2)g. \text{ (5)}$$

Using the gamma function, the correction coefficient J is defined as

$$J(m) = \frac{\Gamma(m/2)}{\sqrt{m/2}\,\Gamma\{(m-1)/2\}}. \text{ (6)}$$

The effect sizes g (2) and d (5) are used in various fields of science, but they assume homoscedasticity just like Student's t-test (Student 1908; Fisher 1925). When this assumption of homoscedasticity is violated, Grissom (2001) recommended the use of Glass's Δ (4) instead of d (5). However, Glass's Δ (4) and d (5) have different meaning because of the difference in denominator. Therefore, Glass's Δ (4) cannot substitute for d (5) in a strict sense. Behavior of g (2), Δ (4), and d (5) under heteroscedasticity was studied in Marfo & Okyere (2019), although the justification for using effect size parameter α, that they defined, to measure the statistic bias under heteroscedasticity was not shown.

Bonett (2008) in psychology proposed a confidence interval (CI) of effect size which does not assume homoscedasticity. First, he defined a general effect size estimator

$$\hat{\delta} = \sum_{j=1}^{k} c_j \bar{Y}_j / s, (7)$$

where $\sum_{j=1}^{k} c_j = 0$, $\bar{Y}_j$ is a sample mean, and $s = \sqrt{k^{-1} \sum_{j=1}^{k} s_j^2}$. Concerning effect size of the difference between two means, substituting $k = 2$, $c_1 = 1$ and $c_2 = -1$ gives

$$\hat{\delta} = \frac{\bar{Y}^1 - \bar{Y}^2}{\sqrt{(s_1^2 + s_2^2)/2}}. (8)$$

Then, he assumed its corresponding parameter and its CI. The CI was calculated using approximation of CI (Viectbauer 2007) and variance of the estimator which was approximately calculated without assuming homoscedasticity. The parameters estimated by $\hat{\delta}$ (7) or (8) were not formulated. He defined the CI for heteroscedasticity without defining a parameter, and this can be a problem. When the estimator does not always correspond to a single parameter, the CI of an undefined parameter loses its consistency in what to estimate, and heteroscedasticity or difference of sample sizes can change the correspondence between estimator and a parameter (See section 3.5.2). Although his CI was effective relative to the other CIs in his simulation experiment where the parameter was given a value, what the value meant could change depending on the variance and sample size, and the change could not be expected since the parameter was not formulated.

It should be noted, Cohen (1969) also defined a parameter of an effect size of the difference between a mean and a constant for a normally distributed population $N_1(\mu, \sigma_1^2)$ and a known constant C as

$$\gamma = (\mu - C)/\sigma_1, (9)$$

Cohen (1969) originally referred to this as d_3', but we refer to this as γ (9) to clearly distinguish it from d (5). Cohen (1969) also defined a biased estimator of an effect size for a normally distributed population with the sample value Y_i^1 $(i = 1, \dots, n_1)$, the sample mean $\bar{Y}^1$, and a known constant C as

$$c^{biased} = (\bar{Y}^1 - C)/s_1. (10)$$

The s_1 is the square root of (3). Cohen (1969) originally referred to this as d_s', but we refer this to c^{biased} for the reason described above. To the best of my knowledge, the

unbiased estimator of γ (9) has not been shown.

There are other effect sizes of the difference that do not assume normality or independence. Since their assumption is different from that of effect size we focus on, we do not treat them in detail but briefly introduce them. Dunlap et al. (1996) invented effect size of the difference between two correlated paired groups. Algina et al. (2005) proposed robust effect size of the difference, which is based on *g* (2) using 20% trimmed mean and 20% Winsorized variance assuming that samples are taken from an observing population and another contaminating population.

3.2 Theory

3.2.1. An effect size of the difference between means without assuming homoscedasticity

First, we define the parameter of an effect size of the difference between means for two independently and normally distributed populations $N_1(\mu_1, \sigma_1^2)$ and $N_2(\mu_2, \sigma_2^2)$ as

$$\epsilon_r = \frac{\mu_1 - \mu_2}{\sqrt{(\sigma_1^2 + r\sigma_2^2)/(r+1)}}, \quad (11)$$

where r is a non-negative real number. This parameter is not generalization of δ (1) and is different from it. Then, suppose two independently and normally distributed populations with the samples Y_i^1 $(i = 1, \dots, n_1)$ and Y_i^2 $(i = 1, \dots, n_2)$, and the sample mean $\bar{Y}^1$ and $\bar{Y}^2$. Based on the statistic t_w, the so-called Welch's t (Welch 1938; Welch 1947), a biased estimator of ϵ_r (11) is defined as

$$e^{biased} = t_w/\sqrt{\tilde{n}}, \quad (12)$$

where

$$t_w = \frac{\bar{Y}^1 - \bar{Y}^2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}, \quad (13)$$

s_i^2 is the same as (3), and

$$\tilde{n} = n_1 n_2/(n_1 + n_2). \quad (14)$$

Finally, e, the unbiased estimator of ϵ_r (11), is

$$e = e^{biased} J(f). \quad (15)$$

Therefore,

$$\mathrm{E}(e) = \epsilon_r.$$

Here, r corresponds to the ratio n_1/n_2. J is the correction coefficient that is defined in equation (6). The degree of freedom *f* is approximately calculated using the

Welch-Satterthwaite equation (Welch 1938; Satterthwaite 1941) as

$$f = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{s_1^4/\{n_1^2(n_1-1)\} + s_2^4/\{n_2^2(n_2-1)\}}. \quad (16)$$

The variance of e (15) is

$$\mathrm{var}(e) = f/(f-2)J^2(f)(1/\tilde{n} + \epsilon_r^2) - \epsilon_r^2.$$

Although this effect size is derived from the difference, we refer to it as e not d. This is because Cohen's d (2) and Hedges' d (5) already exist, and more d would cause further confusion. The proof of the bias correction and variance derivation does not assume homoscedasticity (see the Appendix). In addition, e (15) is a consistent estimator of ϵ_r (11) at the same time. See the Appendix for the proof of the consistency.

3.2.2. An effect size of the difference between a mean and a known constant

Using c^{biased} (10), the unbiased estimator of the effect size parameter γ (9) is defined for a normally distributed population with the sample value Y_i^1 $(i = 1, \ldots, n_1)$, the sample mean $\bar{Y}^1$, and a known constant C as

$$c = c^{biased} J(n_1 - 1). \quad (17)$$

Therefore,

$$\mathrm{E}(c) = \gamma.$$

The correction coefficient J (6) is the same as the one used above. The variance of c is

$$\mathrm{var}(c) = \frac{n_1-1}{n_1-3} J^2(n_1-1)\left(\frac{1}{n_1-1} + \gamma^2\right) - \gamma^2.$$

See the Appendix for proofs of the bias correction and the derivation of the variance. In addition, c (17) is a consistent estimator of γ (9) (See the Appendix for the proof). When interested in constants rather than variables, c' defined as c

$$c' = (C - \bar{Y}^1)J(n_1 - 1)/s_1$$

can be used instead of c.

3.2.3. Confidence intervals of effect sizes

In terms of the effect sizes of the difference, the confidence interval (CI) based on a noncentral t variate is not directly given by a formula (Cumming & Finch 2001). The CI is derived from that of noncentral parameters of noncentral t-distribution, which is in turn obtained by some searching method. The CI based on the biased effect sizes are given as:

$$[ncp_L/\sqrt{\tilde{n}},\ ncp_H/\sqrt{\tilde{n}}]: CI\ based\ on\ g\ and\ e^{biased},$$

and

$$[ncp_L/\sqrt{n_1-1},\ ncp_H/\sqrt{n_1-1}]: CI\ based\ on\ c^{biased},$$

where ncp_L is the noncentral parameter that gives the upper limit of cumulative probability (e.g., 0.975 cumulative probability for 95 % CI) for noncentral t-distribution with the corresponding t value (see the discussion section) and the degree of freedom, and ncp_H is that which gives the lower limit (e.g., 0.025 cumulative probability for 95 % CI), and $\tilde{n}$ and n_1 are the same as (14) and (10). The CIs based on the unbiased estimator of the effect sizes are given by multiplying the corresponding correction coefficient J (6) of the corresponding degree of freedom to the above intervals.

The CI by Bonett (2008) is calculated using variance of the estimator which is approximately calculated without assuming homoscedasticity and approximate assumption of CI (Viechtbauer 2007). Therefore, it is not necessary to apply Bonett's CI to e (15) or c (17), because the derivation of their CIs does not assume homoscedasticity, and their exact CIs can be calculated without approximation.

3.3 Calculation Method

We developed a new package es.dif for R (R core team 2019). It enables the statistics d (5), e (15), c (17), their biased statistics, variance, and CI based on the two samples or their mean, variance, and sample size to be computed. In this package, approximation of J (6) (Hedges 1981) is not employed unless its degree of freedom exceeds 342, when the gamma function returns values that are too large to be treated in R. The CI is obtained by binary search. The figure in this article was drawn using this package.

The remainder of this section presents some examples of the package. First, the following script calculates d (5), e (15), their variances and 95% CIs for data 1 (0,1,2,3,4) and data 2 (0,0,1,2,2).

```
> library(es.dif)
> data1<-c(0,1,2,3,4)
> data2<-c(0,0,1,2,2)
> es.d(data1,data2)
[,1] [,2]
[1,] "Hedges' d:" "0.682379579593354"
  [2,] "variance:" "0.484026380702367"
[3,] "CI:" "[ -0.503527216375147 , 1.82938058482178 ]"
> es.e(data1,data2)
```

```
[,1] [,2]
[1,] "Unbiased e:" "0.668264936033828"
[2,] "variance:" "0.506830833214916"
[3,] "CI:" "[ -0.50334965496395 , 1.7965317007171 ]"
```

Using options of the function, one can change the type I error rate for the CI, calculate biased effect sizes, and output results in the vector style. For example, c^{biased} (10) with 99% CI in the vector style is calculated by this script.

```
> library(es.dif)
> data1<-c(0,0,1,2,2)
> data2<-c(2)
> es.c(data1,data2,alpha=0.01, unbiased=FALSE,vector_out=TRUE)
[1] -1.0000000 0.9292037 -2.5390625 0.5778885
```

In the vector-style output, the four values in the vector show the effect size, its variance, and lower and higher limits of the CI. In addition, this package includes functions that can output effect sizes from the (estimated) parameters and the sample sizes. The following scripts compute d (5) and e (15) for two populations, $N(1,2)$ and $N(0,1)$ with the sample size 5 and 10, respectively.

```
> library(es.dif)
> mean1<-1
> mean2<-0
> var1<-2
> var2<-1
> n1<-5
> n2<-10
> es.para.d(mean1,mean2,var1,var2,n1,n2)
[,1] [,2]
[1,] "Hedges' d:" "0.82286529714397"
[2,] "variance:" "0.349443397657368"
[3,] "CI:" "[ -0.248827687382689 , 1.86616833367494 ]"
> es.para.e(mean1,mean2,var1,var2,n1,n2)
[,1] [,2]
[1,] "Unbiased e:" "0.674259756444758"
```

[2,] "variance:" "0.41613476136966"
[3,] "CI:" "[-0.354146439977423 , 1.65626025590509]"

These types of functions also have the options for the type I error rate, the biased effect size, and the vector-style output.

3.4 Application & Simulation

3.4.1. Practical application

While the situation to use c (17) is clearly different, the e (15) and d (5) have a similar application range in practice. Therefore, we prepared an example of the applications in which the sample variances are not equal. Table 3.1 shows well-known data of three *Iris* species by Fisher (1936), which can also be checked in R (R core team 2019) using a command “iris". Note that only the petal width of *I. setosa* has fewer significant digits. For this data, we calculated d (5), e (15), the ratio of d (5) to e (15), and the ratio of the standard deviations of the two comparing data. Theoretically, e (15) is a more precise estimator of its own parameter than d (5) under this heteroscedasticity.

The calculated result is shown in Table 3.2. When considering their significant digits, the comparing pair of the sepal length of *I. setosa* and *I. virginica* showed the different effect size of d (5) and e (15) (in bold in Table 3.2). Even though most pairs showed identical values of d (5) and e (15), the result revealed that violation of the assumption of homoscedasticity in d (5) can affect the result even in two significant digits.

Figure 3.1 shows the ratio of d (5) to e (15) plotted against the ratio of standard deviations of the comparing data. This figure shows that the similar two standard deviations give similar d (5) and e (15). In other words, the more different two standard deviations are, the more the use of e (15) over d (5) is encouraged.

3.4.2. Simulation

To examine the nature of d (5) and e (15), I also conducted a simulation study. In addition to d (5) and e (15), Bonett’s statistic $\hat{\delta}$ (8) was also included as a reference, although its accuracy cannot be discussed because of the lack of the parameter definition. The above effect sizes and their width of 95% CI were calculated for 100,000 Monte Calro replications from $N(1, \sigma_1^2)$ and $N(0, \sigma_2^2)$ for each condition, and they were represented by their average values. The population means were fixed to 1 and 0. The sample sizes were changed from 10 to 30 in steps of 10. The population standard

deviation σ_1 was fixed to 1 and σ_2 was changed from 1 to 10 in steps of 1. However, some redundant data were omitted from the result. The calculation was conducted using es.dif R package shown above and metafor R package (Viechtbauer 2010). The R source code used for the simulation was shown in the Appendix.

Table 3.3 shows the result of the simulation. When the sample size ratio was conserved under $\sigma_1 \neq \sigma_2$, e (15) gave more similar and concordant values than d (5). For example, e (15) for $n_1 = n_2 = 10, 20, 30$ under $\sigma_2 = 10$ were 0.142, 0.140 and 0.141, whereas the corresponding d (5) were 0.148, 0.143 and 0.143. This is the nature and advantage of e (15) which is designed to estimate the same parameter under heteroscedasticity and the same sample size ratio. The width of CI was narrowest for d (5) under $\sigma_1 = \sigma_2$, and the second e (15) had the second narrowest. Under $\sigma_1 \neq \sigma_2$, e (15), e (15) and $\hat{\delta}$ (8) had the narrowest CI under $n_1 = n_2$, $n_1 > n_2$ and $n_1 < n_2$, respectively. The narrowest CIs of e (15) were followed by d (5), whereas what followed the narrowest CIs of $\hat{\delta}$ (8) was not fixed. It was shown that e (15) had wider situation under which it had the narrowest or second narrowest CI than d (5) or $\hat{\delta}$ (8). Bonett's statistic $\hat{\delta}$ (8) equaled to d (5) under $n_1 = n_2$ as their definition. Under $n_1 \neq n_2$ and $\sigma_1 \neq \sigma_2$, e (15) was closer to $\hat{\delta}$ (8) than d (5). This might imply relative accuracy of $\hat{\delta}$ (8) over d (5) under heteroscedasticity.

3.5 Discussion

3.5.1. Correspondence of effect sizes and *t* tests

Comparison of *t* tests and the effect sizes of the difference except $\hat{\delta}$ (8) shows the clear correspondence between them (Table 3.4). Statistic d (5) corresponds to the unpaired two-sample *t* test (Student 1908; Fisher 1925), whose statistic is the basis of g (2). Statistic e^{biased} (12) uses the statistic (13) of Welch's *t* test (Welch 1947), which aims to test two means with unequal variances, and c^{biased} (10) uses the same statistic as the one-sample *t* test (Fisher 1925). Considering this, it is natural that power analyses should be conducted, using the corresponding pair of the effect size and *t* test. In other words, power analyses of Student's one-sample *t* test, Student's unpaired two-sample *t* test, and Welch's *t* test should be conducted based on the c statistic (17), d (5), and the e statistic (15), respectively. Co-use of non-corresponding t-test and effect size causes inconsistence of the assumption about the population(s).

3.5.2. Influence of sample size on effect size

In this subsection, the relationship between the effect sizes of the difference and sample sizes is described. The value of g (2), a biased estimator of the effect size of the

difference under homoscedasticity, is independent of the sample sizes when the assumption of homoscedasticity ($s_1 = s_2$) is fulfilled. When $s_1 \neq s_2$, it depends on the ratio $q = (n_1 - 1)/(n_2 - 1)$ as implied in (Grissom 2000). This is because g (2) is no longer an estimator of δ (1) under $s_1 \neq s_2$, and it will be a biased estimator of the other parameter δ'_q, which is

$$\delta'_q = \frac{\mu_1 - \mu_2}{\sqrt{(q\sigma_1^2 + \sigma_2^2)/(1+q)}}.$$

Note that even d (5) cannot be the unbiased estimator of δ'_q when $s_1 \neq s_2$, because g (2) is not distributed as non-central *t* variate in this situation. Even if n_1 and n_2 vary, g (2) roughly estimates the same parameter, given the ratio q is fixed.

Next, the e^{biased} (12) is a biased estimator of ϵ_r (11), but ϵ_r (11) equals to the other parameters in the particular situation. When $s_1 = s_2$, $\epsilon_r = \delta$, and e^{biased} (12) equals to g (2), and is independent of the sample sizes. When $s_1 \neq s_2$ and $n_1 = n_2$, $\epsilon_r = \delta'_q$. In this case, e^{biased} (12) equals to g (2) and is also independent of the sample sizes. While d (5) is a biased estimator of δ'_q, e (15) is its unbiased estimator. Therefore, usage of e (15) is always preferable to d (5) in this situation. When $s_1 \neq s_2$ and $n_1 \neq n_2$, e^{biased} (12) depends on the rate $r = n_1/n_2$. Therefore, strictly speaking, multiple e^{biased}s can be comparable only when the sample size ratio r is identical.

The effect size estimator $\hat{\delta}$ (8) did not had a defined parameter, but when $n_1 = n_2$ and $s_1 = s_2$, $\hat{\delta}$ (8) equals to g (2) and e^{biased} (12), and is independent of sample size. Under $n_1 = n_2$ and $s_1 \neq s_2$, $\hat{\delta}$ (8) also equals to g (2) and suffers from the same problem as it. Under $s_1 \neq s_2$ and $n_1 \neq n_2$, the value of $\hat{\delta}$ (8) is no longer the same as g (2), and precise discussion on its behavior is hindered by the lack of its parameter definition. When trying to consider $\hat{\delta}$ (8) as a noncentral t-variate like the other effect sizes, its degree of freedom should be about $n_1 + n_2 - 2$, and n_1 and n_2 should affect the degree of freedom under $s_1 \neq s_2$.

Unlike g (2) or e^{biased} (12), c^{biased} (10) is always independent of the sample size.

The behavior of the unbiased estimator of the effect sizes (d (5), e (15), and c (17)) are almost identical to those that are biased, but they slightly increase as the sample sizes become large. This is because of the correction coefficient J (6), and its behavior is illustrated in detail in (Hedges 1981).

For t tests, there can be a discussion on meaning of testing two means with different variances. Some people might think that such test is meaningless because these tests aim to check whether two groups are equal or not and different variances

themselves mean that they are different groups. Meanwhile, it is said that when one tries to test the central tendency of two groups, such tests are meaningful and that tests which do not assume variance equality should be employed (Ruxton 2006). The same discussion is applicable to effect sizes of the difference; when one wants to check the central tendencies of two groups rather than the whole distribution of the two, the e (15) can provide the difference of two central tendency in terms of the standard deviations. When $r = n_1/n_2 = 1$, for example, the difference of two means are expressed in units of the average of the two standard deviations. As shown above, the value of e (15) equals to that of d (5) under homoscedasticity, and e (15) can be unbiased statistics even under heteroscedasticity. Plus, the above simulation showed that e (15) has wider situations under which its CI is narrower than $\hat{\delta}$ (8). Therefore, in terms of the effect size of the difference between two means, usage of e (15) is preferable to d (5) or $\hat{\delta}$ (8), and e (15) can be the remedy for application of effect size of the difference under heteroscedasticity. However, when the ratio of the two sample sizes cannot be set as uniform under heteroscedasticity, neither d (5) nor e (15) can be precisely compared. This is a form of the Behrens-Fisher problem, which cannot be solved strictly.

3.5.3. Potential applications of the new effect sizes

The effect size e (15) has a vast applicable range covering all kinds of natural and social sciences. This is because e (15) corresponds to Welch's *t* test, whose use is nowadays encouraged over Student's *t* test (e.g., Ruxton 2006). The effect size e (15) is the best option, especially when the ratio of the sample sizes of two groups can be fixed. The effect size c (17) has a relatively narrower range regarding the application. In comparison of paired two groups (the difference in pairs vs. 0) and in some simulation studies (result of simulation vs. the optimal value) or physics (result of experiment vs. physical constant), an effect size of the constant may be needed.

3.6 Figures and Tables

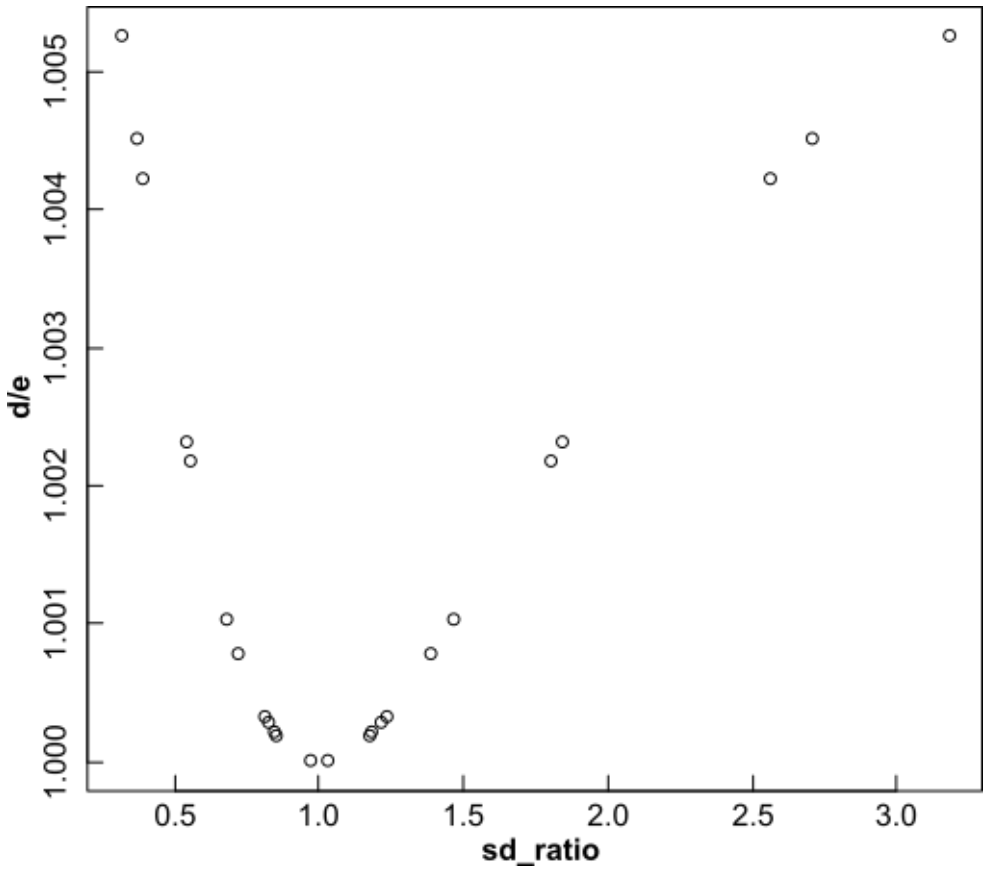

Figure 3.1. Plotted graph of Table 3.2.

Table 3.1. Measured characteristics (in centimeters) of three Iris species shown in Fisher (1936).

Iris setosa				***Iris versicolor***				***Iris virginica***			
S.L.	**S.W.**	**P.L.**	**P.W.**	**S.L.**	**S.W.**	**P.L.**	**P.W.**	**S.L.**	**S.W.**	**P.L.**	**P.W.**
5.1	3.5	1.4	0.2	7.0	3.2	4.7	1.4	6.3	3.3	6.0	2.5
4.9	3.0	1.4	0.2	6.4	3.2	4.5	1.5	5.8	2.7	5.1	1.9
4.7	3.2	1.3	0.2	6.9	3.1	4.9	1.5	7.1	3.0	5.9	2.1
4.6	3.1	1.5	0.2	5.5	2.3	4.0	1.3	6.3	2.9	5.6	1.8
5.0	3.6	1.4	0.2	6.5	2.8	4.6	1.5	6.5	3.0	5.8	2.2
5.4	3.9	1.7	0.4	5.7	2.8	4.5	1.3	7.6	3.0	6.6	2.1
4.6	3.4	1.4	0.3	6.3	3.3	4.7	1.6	4.9	2.5	4.5	1.7
5.0	3.4	1.5	0.2	4.9	2.4	3.3	1.0	7.3	2.9	6.3	1.8
4.4	2.9	1.4	0.2	6.6	2.9	4.6	1.3	6.7	2.5	5.8	1.8
4.9	3.1	1.5	0.1	5.2	2.7	3.9	1.4	7.2	3.6	6.1	2.5
5.4	3.7	1.5	0.2	5.0	2.0	3.5	1.0	6.5	3.2	5.1	2.0
4.8	3.4	1.6	0.2	5.9	3.0	4.2	1.5	6.4	2.7	5.3	1.9
4.8	3.0	1.4	0.1	6.0	2.2	4.0	1.0	6.8	3.0	5.5	2.1
4.3	3.0	1.1	0.1	6.1	2.9	4.7	1.4	5.7	2.5	5.0	2.0
5.8	4.0	1.2	0.2	5.6	2.9	3.6	1.3	5.8	2.8	5.1	2.4
5.7	4.4	1.5	0.4	6.7	3.1	4.4	1.4	6.4	3.2	5.3	2.3
5.4	3.9	1.3	0.4	5.6	3.0	4.5	1.5	6.5	3.0	5.5	1.8
5.1	3.5	1.4	0.3	5.8	2.7	4.1	1.0	7.7	3.8	6.7	2.2
5.7	3.8	1.7	0.3	6.2	2.2	4.5	1.5	7.7	2.6	6.9	2.3
5.1	3.8	1.5	0.3	5.6	2.5	3.9	1.1	6.0	2.2	5.0	1.5
5.4	3.4	1.7	0.2	5.9	3.2	4.8	1.8	6.9	3.2	5.7	2.3
5.1	3.7	1.5	0.4	6.1	2.8	4.0	1.3	5.6	2.8	4.9	2.0
4.6	3.6	1.0	0.2	6.3	2.5	4.9	1.5	7.7	2.8	6.7	2.0
5.1	3.3	1.7	0.5	6.1	2.8	4.7	1.2	6.3	2.7	4.9	1.8
4.8	3.4	1.9	0.2	6.4	2.9	4.3	1.3	6.7	3.3	5.7	2.1
5.0	3.0	1.6	0.2	6.6	3.0	4.4	1.4	7.2	3.2	6.0	1.8
5.0	3.4	1.6	0.4	6.8	2.8	4.8	1.4	6.2	2.8	4.8	1.8
5.2	3.5	1.5	0.2	6.7	3.0	5.0	1.7	6.1	3.0	4.9	1.8
5.2	3.4	1.4	0.2	6.0	2.9	4.5	1.5	6.4	2.8	5.6	2.1
4.7	3.2	1.6	0.2	5.7	2.6	3.5	1.0	7.2	3.0	5.8	1.6
4.8	3.1	1.6	0.2	5.5	2.4	3.8	1.1	7.4	2.8	6.1	1.9
5.4	3.4	1.5	0.4	5.5	2.4	3.7	1.0	7.9	3.8	6.4	2.0

5.2	4.1	1.5	0.1	5.8	2.7	3.9	1.2	6.4	2.8	5.6	2.2
5.5	4.2	1.4	0.2	6.0	2.7	5.1	1.6	6.3	2.8	5.1	1.5
4.9	3.1	1.5	0.1	5.4	3.0	4.5	1.5	6.1	2.6	5.6	1.4
5.0	3.2	1.2	0.2	6.0	3.4	4.5	1.6	7.7	3.0	6.1	2.3
5.5	3.5	1.3	0.2	6.7	3.1	4.7	1.5	6.3	3.4	5.6	2.4
4.9	3.1	1.5	0.1	6.3	2.3	4.4	1.3	6.4	3.1	5.5	1.8
4.4	3.0	1.3	0.2	5.6	3.0	4.1	1.3	6.0	3.0	4.8	1.8
5.1	3.4	1.5	0.2	5.5	2.5	4.0	1.3	6.9	3.1	5.4	2.1
5.0	3.5	1.3	0.3	5.5	2.6	4.4	1.2	6.7	3.1	5.6	2.4
4.5	2.3	1.3	0.3	6.1	3.0	4.6	1.4	6.9	3.1	5.1	2.3
4.4	3.2	1.3	0.2	5.8	2.6	4.0	1.2	5.8	2.7	5.1	1.9
5.0	3.5	1.6	0.6	5.0	2.3	3.3	1.0	6.8	3.2	5.9	2.3
5.1	3.8	1.9	0.4	5.6	2.7	4.2	1.3	6.7	3.3	5.7	2.5
4.8	3.0	1.4	0.3	5.7	3.0	4.2	1.2	6.7	3.0	5.2	2.3
5.1	3.8	1.6	0.2	5.7	2.9	4.2	1.3	6.3	2.5	5.0	1.9
4.6	3.2	1.4	0.2	6.2	2.9	4.3	1.3	6.5	3.0	5.2	2.0
5.3	3.7	1.5	0.2	5.1	2.5	3.0	1.1	6.2	3.4	5.4	2.3
5.0	3.3	1.4	0.2	5.7	2.8	4.1	1.3	5.9	3.0	5.1	1.8
5.0	3.4	1.5	0.2	5.9	2.8	4.3	1.3	6.6	3.0	5.6	2.0
0.35	0.38	0.17	0.1	0.52	0.31	0.47	0.20	0.64	0.32	0.55	0.27

Note: S.L. = sepal length; S.W. = sepal width; P.L. = petal length; P.W. = petal width. The last two rows show the average and the standard deviation of the corresponding column.

Table 3.2. Calculated effect sizes of the difference for the data shown in Table 3.1.

Chara.	Taxa	d	e	d/e	sd ratio
S.L.	1 vs 2	-2.1	-2.1	1.001029	0.682893
	1 vs 3	**-3.1**	**-3.0**	1.002185	0.554334
	2 vs 3	-1.1	-1.1	1.000328	0.811744
S.W.	1 vs 2	1.8	1.8	1.000285	1.214233
	1 vs 3	1.2	1.2	1.000212	1.181483
	2 vs 3	-0.64	-0.64	1.000006	0.973028
P.L.	1 vs 2	-7.8	-7.8	1.004510	0.369243
	1 vs 3	-9.9	-9.9	1.005256	0.314392
	2 vs 3	-2.5	-2.5	1.000197	0.851450
P.W.	1 vs 2	-7	-7	1.002318	0.542139
	1 vs 3	-8	-8	1.004222	0.390349
	2 vs 3	-2.9	-2.9	1.000781	0.720017

Note: Chara. = characteristics; S.L. = sepal length; S.W. = sepal width; P.L. = petal length; P.W. = petal width; Taxa = compared taxa; 1 = *I. setosa*; 2 = *I. versicolor*; 3 = *I. virginica*; d = effect size d (5); e = effect size e (15). These effect sizes are shown in the original significant digits. d/e = the ratio of d (5) to e (15); sd ratio = the ratio of the standard deviations of the compared data. Note that reverse comparisons, such as 2 vs 1, were also conducted, but omitted from this table because their effect sizes are the opposites of the original values, and d/e and sd ratio are the inverses of the original ones.

Table 3.3. Comparison of effect sizes in simulation.

n_1	n_2	σ_1	σ_2	d.ES	d.Par.	e.ES	e.Par.	B.ES	B.Par.	d.CI	e.CI	B.CI
10	10	1	1	1.000	1.000	0.995	1.000	1.000	U.D.	1.823	1.828	1.911
10	10	1	4	0.355	N.C.	0.344	0.343	0.355	U.D.	1.722	1.691	1.834
10	10	1	7	0.210	N.C.	0.201	0.200	0.210	U.D.	1.713	1.668	1.825
10	10	1	10	0.148	N.C.	0.142	0.141	0.148	U.D.	1.710	1.661	1.822
10	20	1	1	0.998	1.000	0.997	1.000	1.002	U.D.	1.579	1.604	1.646
10	20	1	4	0.303	N.C.	0.408	0.408	0.346	U.D.	1.493	1.503	1.310
10	20	1	7	0.175	N.C.	0.243	0.243	0.203	U.D.	1.487	1.488	1.274
10	20	1	10	0.124	N.C.	0.173	0.171	0.143	U.D.	1.486	1.483	1.264
10	30	1	1	0.999	1.000	1.000	1.000	1.007	U.D.	1.482	1.536	1.551
10	30	1	4	0.284	N.C.	0.458	0.459	0.343	U.D.	1.412	1.427	1.096
10	30	1	7	0.164	N.C.	0.278	0.277	0.202	U.D.	1.408	1.416	1.046
10	30	1	10	0.114	N.C.	0.196	0.197	0.141	U.D.	1.407	1.411	1.032
20	10	1	1	1.002	1.000	1.001	1.000	1.006	U.D.	1.580	1.605	1.647
20	10	1	4	0.431	N.C.	0.302	0.302	0.360	U.D.	1.523	1.462	1.830
20	10	1	7	0.262	N.C.	0.175	0.174	0.213	U.D.	1.515	1.443	1.843
20	10	1	10	0.186	N.C.	0.123	0.122	0.151	U.D.	1.512	1.438	1.846
20	20	1	1	0.999	1.000	0.998	1.000	0.999	U.D.	1.303	1.305	1.333
20	20	1	4	0.347	N.C.	0.342	0.343	0.347	U.D.	1.234	1.227	1.277
20	20	1	7	0.204	N.C.	0.200	0.200	0.204	U.D.	1.227	1.214	1.268
20	20	1	10	0.143	N.C.	0.140	0.141	0.143	U.D.	1.225	1.211	1.266
20	30	1	1	1.001	1.000	1.000	1.000	1.001	U.D.	1.189	1.199	1.215
20	30	1	4	0.318	N.C.	0.379	0.378	0.346	U.D.	1.125	1.129	1.050
20	30	1	7	0.184	N.C.	0.222	0.222	0.201	U.D.	1.120	1.119	1.032
20	30	1	10	0.130	N.C.	0.157	0.157	0.142	U.D.	1.119	1.115	1.027
30	10	1	1	0.998	1.000	0.999	1.000	1.005	U.D.	1.482	1.536	1.551
30	10	1	4	0.486	N.C.	0.285	0.286	0.362	U.D.	1.447	1.378	1.833
30	10	1	7	0.304	N.C.	0.166	0.164	0.216	U.D.	1.442	1.360	1.854
30	10	1	10	0.212	N.C.	0.113	0.115	0.148	U.D.	1.440	1.355	1.859
30	20	1	1	0.999	1.000	0.999	1.000	1.000	U.D.	1.189	1.199	1.215
30	20	1	4	0.386	N.C.	0.316	0.316	0.349	U.D.	1.133	1.120	1.268
30	20	1	7	0.229	N.C.	0.184	0.183	0.205	U.D.	1.127	1.108	1.269
30	20	1	10	0.161	N.C.	0.129	0.129	0.144	U.D.	1.125	1.105	1.269
30	30	1	1	1.002	1.000	1.001	1.000	1.002	U.D.	1.067	1.069	1.084

30	30	1	4	0.347	N.C.	0.343	0.343	0.347	U.D.	1.011	1.010	1.037
30	30	1	7	0.204	N.C.	0.202	0.200	0.204	U.D.	1.006	1.000	1.029
30	30	1	10	0.143	N.C.	0.141	0.141	0.143	U.D.	1.005	0.998	1.027

Note: d = effect size d (5); e = effect size e (15); B = effect size $\hat{\delta}$ (8); Par. = parameter of effect size; CI = width of confidence interval; N.C. = not calculable; U.D. = undefined. The narrowest CI in each row is underlined.

Table 3.4. Correspondence of assumptions, t values, and effect sizes of the difference.

	One sample & a constant	Two samples under homoscedasticity	Two samples under heteroscedasticity
As.	Normality	Normality, Independence, & Homoscedasticity	Normality & Independence
t	$t = \frac{\bar{Y}^1 - C}{\sqrt{s_1^2/(n_1 - 1)}}$	$t = \frac{\bar{Y}^1 - \bar{Y}^2}{S^{pooled}/\sqrt{\tilde{n}}}$	$t = \frac{\bar{Y}^1 - \bar{Y}^2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$
ES	$c = \frac{\bar{Y}^1 - C}{s^1}J$	$t = \frac{\bar{Y}^1 - \bar{Y}^2}{S^{pooled}}J$	$t = \frac{\bar{Y}^1 - \bar{Y}^2}{\sqrt{(s_1^2/n_1 + s_2^2/n_2)\tilde{n}}}J$

Note: As. = assumption; t = t value; ES = effect size. The degree of freedom of J is omitted for the space and must be calculated corresponding degree of freedom.

Chapter 4. Application of New Sampling Method

Phylogenetic status and taxonomy of *Nanocnide lobata* and *N. pilosa* (Urticaceae)

4.1 Introduction

The genus *Nanocnide* Blume (1856: 155) in Urticaceae is a genus of herbaceous plants distributed in warm-temperate to tropical regions of East and Southeast Asia, and six species have been reported in this genus: *Nanocnide japonica* Blume (1856: 155), *N. lobata* Weddell (1869: 69), *N. closii* H. Léveillé & Vaniot in H. Léveillé (1904: cxliv as "Closii"), *N. dichotoma* S. S. Chien in Pei (1934: 142), *N. pilosa* Migo (1934: 386) and *N. zhejiangensis* X.F. Jin & Y.F. Lu (2019: 5). *Nanocnide japonica*, the type species of this genus, was described from Japan, and is distributed in Japan, Korea, Taiwan, and Mainland China (Jiarui et al. 2003). *Nanocnide lobata* was described from "Loo-Choo", which corresponds to current Ryukyu Islands in Japan, and is distributed in Mainland China, and Vietnam (Jiarui et al. 2003). While Jiarui et al. (2003) wrote this plant grew in Taiwan, there was no record of this plant from Taiwan as far as I searched Herbarium TAI, the field in Taiwan and a literature about Taiwan (Yang et al. 1996). Here, the herbarium codes follow Thiers & Tulig (2019). Also, this plant grows in Ryukyu Islands (including Amami Islands) in Japan as described in the original literature. *Nanocnide closii* described from Mainland China was currently synonymized under two species (Jiarui & Monro 2003; Huaxing & Gilbert 2008). One of them is *Pilea japonica* (Maximovicz) Handel-Mazzetti (1929: 141) in Urticaceae which was originally described as *Achudemia japonica* Maximovicz (1876: 627) from Japan (Jiarui & Morno 2003). Another species which synonymizes *N. closii* is *Acalypha supera* Forsskål (1775: 162) in Euphorbiaceae (Huaxing & Gilbert 2008). I checked the digital image of the holotype of *N. closiii* (*J. Cavalerie 2732* in E as E00185121), and concluded *N. closii* should be a synonym of *A. supera*. I do not treat this species any more in this article. *Nanocnide dichotoma* is synonymized under *N. japonica* (Jiarui et al. 2003; Tateishi 2006). I checked the holotype, and I agree to adopt this treatment. Previous phylogenetic analyses treating *N. japonica* also have not shown the evidence of the cryptic species which may correspond to *N. dichotoma*. I will mention these previous phylogenetic analyses later in detail. *Nanocnide pilosa* was originally described from Mainland China. This species was treated as a synonym of *N. lobata* by Jiarui et al. (2003), while it was treated as a distinct species by Tateishi (2006) (Table 4.1). Although the classification of Jiarui et al. (2003) is widely accepted except in Japan, *N. pilosa* is accepted and treated as an endangered species in Japan. Solution of this taxonomic contradiction is the main purpose of this study, and inspection of the

phylogeography and cryptic taxa is the subsidiary purpose. Finally, *N. zhejiangensis* was recently reported from Zhejiang, China. While *N. zhejiangensis* is morphologically and phylogenetically close to *N. japonica*, *N. zhejiangensis* is a monophyly independent of *N. japonica* and has morphological differences. To solve the taxonomic problem about *N. pilosa* and *N. lobata*, I planned to conduct phylogenetic analyses of this genus and to suggest a classification based on molecular phylogeny. As a working hypothesis, I temporally admit *N. pilosa* and *N. lobata* sensu Tateishi as distinct operational taxonomic units (OTU).

Several studies have reported the molecular phylogeny of *Nanocnide* using a few samples mainly as an outgroup. Wu et al. (2013) studied phylogeny of Urticaceae using nuclear ribosomal internal transcribed spacer (ITS) regions and four chloroplast regions. They showed that four *Nanocnide* samples from Mainland China made a monophyly with full support (100% posterior probability in the Bayesian inference (BI), 100% bootstrap value in the maximum likelihood (ML) tree, and 100% bootstrap value in the maximum parsimony (MP) tree). The clade of *Nanocnide* was located as a sister group of the clade of *Urtica* (Urticaceae). Henning et al. (2014) studied a *Urtica* species, and one *N. japonica* from Mainland China and one *N. lobata sensu* Jiarui et al. from Japan were included in the phylogenetic analysis based on ITS and three chloroplast regions. As a result, they also showed monophyly of *Nanocnide* with full support (BI, ML, and MP). With almost the same members of the former study, Grosse-Veldmann et al. (2016) studied phylogeny of whole *Urtica*, and the same *Nanocnide* samples as the former study were included in the analysis as the outgroup. In their phylogenetic tree based on concatenated sequence of ITS and three chloroplast non-coding regions, *N. japonica* and *N. lobata sensu* Jiarui et al. from Mainland China made a monophyly with full support (BI & ML), and this clade made a monophyly with *Laportea cuspidata* (Weddell) Friis (1981: 156) (Urticaceae) with full support. Kim et al. (2015) studied phylogeny of Urticeae (Urticaceae) including *Nanocnide* and *Urtica* using ITS and two chloroplast regions. They used four *Nanocnide* samples from Mainland China, and also showed monophyly of *Nanocnide* with full support (BI & MP) and monophyly of *Nanocnide* and *L. cuspidata* with (99% bootstrap value in MP and full support in BI). Tseng et al. (2019) analyzed phylogeny of *Elatostema sensu lato* (Urticaceae). One sample of *N. japonica* from Taiwan was included in it, but taxa close to *Nanocnide* were not included in it, considering the tree by Wu et al. (2013). Jin et al. (2019) conducted a phylogenetic analysis based on ITS and three chloroplast regions using *N. lobata sensu* Jiarui et al. from Mainland China, *N. japonica* from Mainland China and Japan, and *N. zhejiangensis* from Mainland China. Their result of the concatenated tree showed

monophylies of the three species and monophyly of the genus *Nanocnide*.

Most *Nanocnide* samples used in previous analyses were those from Mainland China, and relationship among multiple areas were not analyzed. Concerning the problem of *N. lobata* and *N. pilosa*, only the sample sequenced by Henning et al. (2014) (Japan, M. Furuse 2091 K) may correspond to *N. lobata sensu* Tateishi in spite of lack of detailed information in the article. The reason why it may be *N. lobata sensu* Tateishi is that the probable collector, Miyoshi Furuse (1911–1996), had lived in Ishigaki Island in Ryukyu Islands. However, even if this sample corresponded to *N. lobata sensu* Tateishi, there would be only one sample for this taxon, and its monophyly could not be checked with published materials. Phylogenetic analysis with multiple samples is necessary to elucidate phylogenetic and taxonomic status of *N. lobata sensu* Tateishi.

4.2 Materials and Methods

4.2.1. Sampling

Sampling of *Nanocnide* was conducted as follows: Distribution information was collected from herbaria TI, MAK, MBG, HYO, TRPM, and KAG and personal websites. The collected distribution of each taxon was organized in a csv file and processed with Samploc to obtain sampling locations. The number of sampling locations was decided according to my sampling budget and time. The sampling was conducted mainly in Spring, the inflorescent season, when the plants can be easily identified and are rich in characteristics. Therefore, the sampling started from the southern distribution and proceeded northward. In most cases, multiple close sampling locations were visited in a single trip to save money and time. For example, four locations in Kyushu were visited in a row in this study. After every sampling trip, the result of the sampling was reflected to the sample csv file, and sampling locations were re-calculated with Samploc (adaptive sampling; Thompson & Seber 1966). Apart from the newly collected samples, one Taiwanese sample I previously collected was included into the analyses. Plus, one Japanese sample and five samples from Mainland China provided by collaborators were also utilized. Plants other than *Nanocnide* used for the outgroups were sampled when encountered during the field working. The voucher specimens were deposited at TI. In addition to these samples, sequences in GenBank (Table 4.3) were selected and used for the phylogenetic analysis. Total DNA was extracted from silica-dried leaves using DNeasy™ Plant Mini Kit (Qiagen) by following its manual.

4.2.2. Morphological Observation

Some of the flowers of the samples were preserved as freeze-dried specimens to observe them referring the results of the phylogenetic analyses. I cut a raw flower, and put it in a

polystyrene box. Then, I filled the box with silica-gel beads and sealed it with a parafilm®. Finally, I left it in a refrigerator at -30 ℃ for a month at least. I referred to Iino (2007) for this method, but wetting samples before freeze-drying which he suggested was not employed. This is because water pressure damaged structure of flowers when testing this method with *Sonchus oleraceus* Linnaeus (1753: 794) (Asteraceae) and *Persicaria longiseta* (Brujin) Kitaguchi (1937: 322) (Polygonaceae). Observation of morphology was conducted after the phylogenetic analyses using the pressed specimens, their photographs and the freeze-dried specimens.

4.2.3. Phylogenetic Analysis using Sanger Sequencing

The ITS regions were amplified by polymerase chain reaction (PCR) using the primer pair "ITS1" and "ITS4" (White et al. 1990). The PCR conditions were 5 min at 94 °C, followed by 35 cycles of 1 min at 94 °C, 1 min 50 °C annealing, and 1.5 min at 72 °C. The conditions included a final extension step of 10 min at 72 °C. The PCR products were refined with an illustra™ ExoSAP-IT™ (Affimetrix). Using this purified PCR products, sequencing was conducted by Fasmac DNA Sequence Service (Kanagawa, Japan). The resultant bidirectional electropherograms were assembled with GeneStudio™ Pro (http://genestudio.com/), and sequences were scored by the program. Sequences were aligned using the online version of MAFFT version 7 (Katoh et al. 2017) with the default options. Both ends of the aligned sequences were trimmed to the length of the shortest one. Then, completely identical sequences were collapsed and treated as a single OTU.

Two datasets were analyzed: Dataset 1 contained all the samples in Table 4.2 and 4.3 to elucidate phylogeny of the genus *Nanocnide*; dataset 2 was composed of the samples of *Nanocnide* and *L. cuspidata* as the outgroup to inspect phylogeny within *Nanocnide*. The ML and MP analyses were conducted using MEGA6 ver.6.06 (Tamura et al. 2013). The ML analyses were processed with complete deletion, subtree-pruning-regrafting (SPR) level 5, default initial tree, and "Very Weak" branch swap filter options. In the MP analyses, gaps were coded with simple indel coding (Simmons & Ochoterena 2000) using FastGap (Borchsenius 2009). Subsequently, coded data were analyzed with complete deletion, the default number of initial trees, a tree bisection reconnection heuristic tree search, and MP search level 3 options. The BI analyses were conducted using MrBayes 3.2.3. (Ronquist & Huelsenbeck 2003). Markov chain Monte Carlo calculations were repeated until the average standard deviation of split frequencies fell below 0.010. The substitution model for Bayesian analysis was selected based on the AICc4 file calculated with Kakusan4 (Tanabe 2011). The statistical support was assessed by posterior probability on the BI tree, and

bootstrap values on the ML and MP trees with 1000 replicates.

4.2.4. Analyses using Next-Generation Sequencing

Using the DNA extracted in the preceding step, I also conducted phylogenetic analysis with MIG-seq (multiplexed ISSR genotyping by sequencing) to resolve phylogenetic relationship between *N. pilosa* and *N. lobata sensu* Tateishi. All *Nanocnide* samples collected in this study were included in this analysis, and *L. cuspidata* sample collected in this study was also included as the outgroup. The PCR and sequencing were conducted according basically to the original article (Suyama & Matsuki 2015), but the annealing temperature in the first PCR was changed to 38 ℃. MIG-seq data were processed using Stacks (Catchen et al. 2013) and detected loci with default parameters. I called single nucleotide polymorphisms (SNP) using Stacks with r = 0.1, min_maf = 0.01 and max_obs_het = 0.95. Based on these SNPs, ML tree was inferred with using RAxML (Stamatakis 2014). Although I included a *L. cuspidata* as the outgroup, the resultant trees posed it in a group of *N. pilosa* (see Fig. S22 in Appendix). This was considered to be due to long-branch attraction (cf. Bergsten 2005). For this reason, I excluded *L. cuspidata* from the ML analysis, treated *N. japonica* as the outgroup and examined the relationship between *N. pilosa* and *N. lobata*.

To find potential hybridization in *Nanocnide*, Bayesian clustering using Structure v.2.3.4 (Pritchard et al. 2000) was also conducted for all the *Nanocnide* samples. Hybridization can be identified as samples composed of more than one ancestral population which corresponds not to hybrids, but to species (Moe & Weiblen 2012). SNPs were called for this analysis using Stacks with r = 0.7, 0.8 and 0.9, min_maf = 0.01 and max_obs_het = 0.95. I performed this analysis for these three SNP data with the number of parental populations k = 1–5, 100,000 initial burn-in and 100,000 Markov chain Monte Carlo iterations. An admixture model with correlated allele frequency with the default parameters was employed for the analyses. The plausible number of the parental populations k was examined using Structure Harvester web v.0.6.94 (Earl & vonHoldt 2012) As discussed later, the result using all the *Nanocnide* samples assigned a single population to *N. lobata* and *N. pilosa* which made distinct clades in the phylogenetic analyses. To confirm hybridization in *N. lobata* and *N. pilosa*, the analysis using only *N. lobata* and *N. pilosa* samples was also conducted in the same way as described above.

4.3 Results

4.3.1. Sampling

In total, 37 samples were collected, covering the distribution in Japan, Taiwan, and

Southeastern part of Mainland China (Table 4.2). This included one *L. cuspidata* sample and two *Urtica thunbergiana* Siebold & Zuccarini (1846: 214) samples as the outgroup, but *N. zhejiangensis* was not sampled which was reported after finishing this sampling scheme.

4.3.2. Phylogeny inferred by Sanger Sequence

The ML phylogenetic trees of ITS were shown in Figs.4.1 and 4.2. Hereafter, I indicate the statistical support of a branch in the ML, MP, and BI trees like (90/90/0.90).

In the dataset 1 (Fig. 4.1), phylogeny of genera was identical among the ML, MP, and BI trees except the genus *Zhengyia* Deng *et al*. (2013). While *Zhengyia* were sister to the genus *Urtica* in the ML and BI trees with low support (62/0.86), it was sister to *Nanocnide* plus *Laportea cuspidata* clade with low support (46) in MP analysis. In the MP analysis, six MP trees were obtained, and their topologies were identical except for four taxa in *Laportea* clade (KF137870, EU003928, KM586392, and KC284942). Three samples of *Girardinia diversifolia* (EU003927, KY425770, and EU003926) and one sample of *L. cuspidata* (EU003928) were considered to be misidentified because of their phylogeny and morphological resemblance in Urticaceae. Concerning *Nanocnide*, all the samples of *Nanocnide* were included in a clade with full support (100/100/1), and this *Nanocnide* clade was sister to the clade of *L. cuspidata* with high support (83/96/1). The other groups closely related to *Nanocnide* were a sample of *Zhengyia* and the clade of *Urtica*. Although relationships among them were not clearly resolved, monophyly of *Nanocnide*, *L. cuspidata*, *Zhengyia* and *Urtica* was highly supported (84/97/1). In addition, the clade of *L. cuspidata* was apart from the clade containing the other taxa of *Laportea*.

In the dataset 2 (Fig. 4.2), topology of the ML, MP, and BI trees were almost identical. Exceptionally, 2 of 10 MP trees contained a subclade in Clade A with low support (21), and the subclade was composed of all OTUs in *N. japonica* except B2 sample. The support of the branch containing KX271353, KC284944, KC284946 and NG7 was not common among 10 MP trees: 43 for four trees and 38 for six trees. Support in Fig. 2 was represented by the most frequent value. Concerning the phylogeny within *Nanocnide*, all the ingroup samples were divided into three groups for discussion: Clade A, paraphyletic Group B, and Clade C. Clade A, Group B and Clade C were exclusively composed of *N. japonica*, *N. pilosa* and *N. lobata sensu* Tateishi, respectively. The ML, MP and BI trees fully supported Clade A (100/100/1) and the clade composed of Group B and Clade C (100/100/1). While monophyly of *N. lobata sensu* Tateishi was highly supported (96/83/1), phylogenetic status of *N. pilosa* was not clearly resolved and remained paraphyly (67/64/0.66).

4.3.3. Phylogeny inferred by Next-Generation Sequence

The obtained reads per sample ranged from 326,494 to 732,016 and the average and the total sum was 547,920 and 18,629,280, respectively. In total, 18,464 SNPs were detected for the phylogenetic analysis. The mean coverage depth per sample was ranged from 15.5 to 28.1 and its average was 20.6. The tree based on MIG-seq data was shown in Fig. 4.3. Three clades A, B and C were observed, and they corresponded to Clade A, Group B and Clade C seen in ITS trees (Fig. 4.2). In Clade A composed of *N. japonica*, samples from Mainland China and Taiwan diverged early and made a monophyly with 97 and 99 bootstrap values, respectively, whereas monophyly of Japanese sample was poorly supported (9). Within Japanese samples, samples from Kyushu Island diverged early (KY25, KY2 and KY29), but the rest of Japanese samples did not show a clear phylogenetic pattern corresponded to sampling locality. Clade B was composed of *N. pilosa*. However, its confidence was low (68). Japanese samples of Clade B made a monophyly with moderate confidence (87) and samples from Mainland China were paraphyly. Clade C comprised *N. lobata sensu* Tateishi, and was fully supported (100). The sample from Ishigaki Island (R2) was a little diverged from the other samples in Clade C. Concerning the relationship among the three clades, Clade C had considerably a long branch.

4.3.4. Structure Analysis

For the analysis using all the *Nanocnide* samples, 184, 17 and 14 genotypes were detected for r = 0.7, 0.8 and 0.9. The likelihood of each setting was listed on Table 4.4. According to this result, the biggest likelihoods for r = 0.7, 0.8 and 0.9 were observed in k = 3, 4 and 4. The result also showed that k = 2 always gave the biggest increment of likelihood (Δk). Therefore, the results of k = 2–4 were visualized as Figs. 4.4–4.6 and discussed, considering the difficulty of estimating k (Pritchard et al. 2000). All the results clearly distinguished the samples of *N. lobata* and *N. pilosa* from the other samples. Therefore, in the population graphs (Figs. 4.4–4.6), this population was manually arranged at the rightmost position and colored purple for clear understanding.

For the analysis using *N. lobata* and *N. pilosa* samples, 587, 354 and 178 genotypes were detected for r = 0.7, 0.8 and 0.9. The likelihood was listed on Table 4.5, and the results of k = 2, 4 and 5 were visualized as Fig. 4.7. All the results clearly distinguished the samples of *N. lobata* from *N. lobata*, and the population which corresponded to *N. lobata* was colored brown in the all graphs in Fig. 4.7. for clear understanding.

4.3.5. Morphological Observation

The freeze-dried specimens could provide preserved morphology of the flowers which the pressed specimen could not. See descriptions in Taxonomic Treatment for the results of morphological observation.

4.4 Discussion

4.4.1. Sampling

First of all, it should be noted that it is impossible to discuss the effectiveness of the spatial sampling of *Nanocnide* comparing to random sampling in terms of their covered genetic diversity. This is because random sampling cannot be conducted when assuming the mother population as all individuals in a taxon, and this is the reason why effectiveness of spatial sampling was discussed using previous sampling data in Chapter 2.

Applying spatial sampling to a real taxonomic study provided some lessons about spatial sampling or the other related methodology. I visited four sampling location in my first sampling trip to Kinki district. However, I could find no samples in this trip. This seemed due to the old distribution information on which the sampling was based. For example, one of the locations I visited in Mie Prefecture was based on a specimen record in 1907 at herbarium MAK. I concluded that older distribution information was less confident. In Japan, this tendency seemed to be prominent near villages or cities where the environment drastically changed in the last century. To solve this old-record problem, I excluded distribution information at villages before ca. 1945 from Samploc calculation by changing their necessity variables in the csv file. This seemed to function; I could find samples in my second Kinki district sampling at the expense of reducing the total number of the sampling locations. Generally speaking, filtering of distribution information is necessary for efficient sampling. Current Samploc system cannot distinguish such filtering from sampling failure; both of them are recorded as “unnecessary” in necessity variables which means exclusion from the calculation. Distinguishing them on the csv file and publishing it as a record of sampling will make sampling procedures more transparent and reproducible. Plus, ideally, filtering of distribution information should be conducted by some clear and reproducible method. Distribution modeling or some kind of Bayesian inference are probably useful for this purpose.

The other lesson is about computation efficiency of spatial sampling. I used distribution information in herbarium MBK (The Kochi Prefectural Makino Botanical Garden), and it provided many recent records from Kochi Prefecture. However, the records from Kochi Prefecture represented more than 10% of all collected distribution

data of *N. japonica*. This geographical concentration of data is redundancy and reduces efficiency of calculation of spatial sampling by meaninglessly expanding the searching space. This phenomenon is not specific to Kochi samples, but it is prominent when the many sample data concentrated on a specific geographical area. I excluded these data in Kochi from calculation after finishing sampling in Shikoku. This procedure should be implemented into Samploc as its function in some way.

There is no guarantee that one can obtain samples at the calculated sampling locations even if one carefully filters the distribution information, and spatial sampling is an interactive process in which sampling scheme adaptively keeps changing in response to the sampling result. When one failed sampling in a location, Samploc provides the alternative sampling location. Therefore, the number of sampling locations at the first calculation should be kept fewer than the maximum number that one's budget and time to visit can afford. In other words, one should reserve one's budget and time to compensate for possible sampling failure. Otherwise, when sampling scheme was conducted in a specific direction, sampling density will be directionally biased. For example, when the sampling scheme proceeds from south to north and there are no spare budget and time, many locations will be sampled in south and few in north because of the compensation of sampling failure. Of course, this problem does not occur when one can always afford money and time for sampling failure.

If one assumes sampling failure occurs in every sampling location at the same probability, not employing adaptive sampling is reasonable. In this tactics, Samploc is employed only at the start of the sampling scheme and do not conduct re-calculation of the sampling locations during the scheme. If the above assumption is fulfilled, sampling location bias due to sampling failure is expected not to occur. However, such assumption seems unrealistic at least for land plants, because biological population at the edges of the distribution is often poor in its size and has high probability of sampling failure due to failure in finding the population or population extinction. This thought is similar to the abundance-center model in biogeography which assumes that "a species' abundance is typically greatest at the centre of its geographical range and uniformly low toward the edges" (Sagarin & Gaines 2002). The present *N. japonica* sampling at the northernmost localities was actually failed (Ishikawa Prefecture in Sea of Japan side and Iwate Prefecture in Pacific Ocean side).

4.4.2. Analyses on the sequences

From the result of dataset 1 (Fig. 4.1), I demonstrated the monophyly of *Nanocnide* using multiple samples per taxon (100/100/1). Additionally, I confirmed that *Nanocnide* was sister to *L. cuspidata* (83/96/1), and that it was independent of the other *Laportea*

species including the type, *L. canadensis* (Linnaeus 1753: 985) Weddell (1854: 181). These results are concordant with those of Kim et al. (2015) and Grosse-Veldmann et al. (2016). Taxonomic status of *L. cuspidata* requires revision using more samples of it and its form, *L. cuspidata* f. *bulbifera* (Kitamura 1967: 207) Fukuoka et Kurosaki (1995: 89).

From the result of dataset 2 (Fig. 4.2), monophyly of *N. japonica* (Clade A) was confirmed. In spite of its relatively wide distribution, only four haplotypes were observed in *N. japonica*, and I did not recognize a lineage which may correspond to *N. dichotoma* or a new cryptic species. Although monophyly of the clade *N. pilosa* (Group B) plus *N. lobata sensu* Tateishi (Clade C) was fully supported, relationship between them was remained unresolved. Group B had two haplotypes, and both of them include Chinese samples and Japanese samples. Clade C also had two haplotypes, and all of the newly collected samples shared a single haplotype, and another haplotype was observed in Furuse's sample.

Concerning the MIG-seq based tree (Fig. 4.3), the geographically sequential divergence in Mainland China, Taiwan and Kyushu in Clade A implies possibility that *N. japonica* speciated in Mainland China and expanded its distribution eastward. The MIG-seq based tree could not confidently confirm the monophyly of *N. pilosa*. Although this study does not focus on phylogeography, the divergence scenario of *N. pilosa* and *N. lobata sensu* Tateishi is supposed as follows: The common ancestor of *N. pilosa* and *N. lobata* was once distributed in the area including current Mainland China, East China Sea, Ryukyu Islands and Kyushu in some glacial period. Then, rise of the sea level and Okinawa Trough isolated some of the plants in Ryukyu arc, and this geographical isolation caused speciation between *N. lobata* and *N. pilosa*. Since the isolation of Ryukyu Islands by Okinawa Trough preceded to that of Kyusyu by East China Sea, *N. pilosa* in Kyusyu is phylogenetically closer to that in Mainland China in spite of the fact that Ryukyu Islands are geographically closer to Mainland China than Kyusyu. Assuming this scenario, it is somewhat unnatural that Taiwan does not have these plants. They might have been extinct in Taiwan, or just remained unfound.

The long branch of Clade C in MIG-seq tree (Fig. 4.3) was interpreted as follows. First, ITS or the other conventional markers are not translated. On the other hand, MIG-seq method gathers SNPs between single sequence repeats on the whole genome. Therefore, mutations obtained by MIG-seq are considered to be closer to "nearly neutral" rather than "neutral" when comparing to the conventional markers. When mutation is neutral in selection, speed of the mutation fixation is independent of population size, whereas nearly neutral theory tells fixation of non-neutral mutations is

faster in smaller population (Ohta 1973). Now, when considering *N. lobata sensu* Tateishi and *N. pilosa*, both of them are annual plants and their generation spans are equal. Their supposed speciation scenario described above implies that population size of *N. lobata sensu* Tateishi is smaller than that of *N. pilosa*, because Ryukyu arc is much narrower than the continental shelf which corresponds to current East China Sea. Therefore, the long branch of Clade C may be due to the smaller population size of *N. lobata sensu* Tateishi relative to *N. pilosa* comprising Clade B, and the long branch was not conspicuous in the ITS tree (Fig. 4.2).

All the results of Structure analyses (Figs. 4.4–4.6) clearly distinguished *N. lobata* and *N. pilosa* from *N. japonica*, and this can be interpreted as absence of hybridization between *N. japonica* and the population composed of *N. lobata* and *N. pilosa*. The samples of *N. japonica* from Taiwan and Mainland China were a little distinguished in the results of r = 0.7. This may be due to the mutations unique to them which made monophylies in the tree (Fig. 4.3) and were excluded in r = 0.8 and 0.9. Samples from Nagano and Mie were a little distinguished for k = 4 when r = 0.8 and 0.9, and result for r = 0.8 and k = 3 had a unique pattern in which *N. japonica* was separated into two populations. However, these are difficult to interpret; phylogeographical analysis focused only on *N. japonica* with multiple samples per locality might explain this pattern. What is interesting is none of the results distinguished *N. lobata* and *N. pilosa*, although they made distinct clades in the phylogenetic tree (Fig. 4.3). This is because most of the mutation sites which distinguished them were not shared with *N. japonica*. In the 18,464 bases of the data for the phylogenetic analysis (r = 0.1), mutation sites among *N. lobata* and *N. pilosa* were 2,999 bases. Excluding sites which were missing in all *N. japonica* samples from these 2,999 sites provides 626 bases. This means about 79% of mutations among *N. lobata* and *N. pilosa* were located at non-shared sites with *N. japonica*. According to the help file of Structure, such sample-specific distribution of missing data violates the assumption of Structure, and therefore the contradiction between the results of the phylogenetic analysis and the Structure analysis should not be biologically interpreted. The Structure analysis of *N. pilosa* and *N. lobata* (Fig. 4.7) confirmed distinction and absence of hybridization between them. Interspecific structure of *N. pilosa* was observed for k = 4 and 5, but it cannot be confidently discussed with the current data and is out of the purpose of this study.

4.5 Taxonomic Treatment

In this taxonomic treatment, I made species correspond to confident clades (cladistic

species concept). Intra-specific taxa were assigned to groups which had minor or unconfident phylogenetic differences accompanied with their morphological differences. The minor phylogenetic difference here means a monophyly with low confidence or a monophyly that renders a species non-monophyletic if they are considered as a species.

According to above standard, I treat *N. pilosa* as a subspecies of *N. lobata*. This is because although *N. pilosa* had phylogenetical and morphological uniqueness, its monophyly was not confident. The ITS sequences of *N. zhejiangensis* were identical to those of *N. japonica*, but monophyly of *N. zhejiangensis* is confidently confirmed in the previous study using chloroplast regions (Jin et al. 2019). Therefore, I kept the current taxonomic status of *N. zhejiangensis*. Simple descriptions below were based on my observation of the images on the original literature (Jin et al. 2019).

Note that the following new nomenclature is not validly published here (International Code of Nomenclature Art. 30.9; Turland et al. 2018).

Key to taxa in *Nanocnide*

A. Perennial; staminate peduncles elongate above highest leaves; stems glabrous or with appressed hairs ..... B.

A. Annual; staminate peduncles not exceeding highest leaves; stems with depressed hairs ..... C.

B. Staminate perianths strigose ..... **1. *N. japonica***

B. Staminate perianths glabrous ..... **2. *N. zhejiangensis***

C. Staminate flowers terminal or axillary; pistillate perianths terete in fruition; largest leaf with 5 teeth ..... **3a. *N. lobata* subsp. *lobata***

C. Staminate flowers axillary; pistillate perianths laminar in fruition; largest leaf with 11–15 teeth ..... **3b. *N. lobata* subsp. *pilosa***

1. *Nanocnide japonica* Blume, Mus. Bot. 2: 155. 1856. *Typus.* Illustration on Mus. Bot. 2: illust. XVII. (holo- digital image!) –Fig. 4.8. and Fig. 4.11.

= *N. dichotoma* S.S. Chien, Contr. Biol. Lab. Chin. Assoc. Advancem. Sci., Sect. Bot. 9(2): 142, f. 16. 1934. *Typus.* Nanking Shir-er-tung, under cliff, 13 April 1928, *R. C. Ching 5144*. (holo- NAS 00297816 digital image!).

Description. Perennial herbs. Stems glabrous or having appressed hairs, erect before flowering, hibernating stolons creeping on ground after flowering. Leaves opposite when young, alternate when matured, flabellate at lower nodes, deltate at higher nodes, pubescent, rarely urent triggering slight pain, dentate, with 11–16 teeth on edge of largest leaf; petioles elongate, grooved on adaxial surfaces. Stipules opposite, ca. 0.1 cm

long, lanceolate to widely ovate, persistent after leaf deciduation. Flowers diclinous. Staminate inflorescence determinate, umbel or dichasium; peduncles elongate above highest leaves; bracts and bractlets sometimes absent, membranous, lanceolate when present; perianths strigose, five, closed, forming turbinate buds when immature, open when matured, upper surfaces often reddish purple, otherwise green; filaments light yellow, translucent; anthers white; pistillode pentagonal, circularly hollowed at center. Pistillate inflorescence determinate umbel, solitary; peduncles almost absent when young, elongate when matured blown to swing off fruit; pedicels very short; perianths having a spike on apex, four, unequal in size: two longer ones lance-ovate, cymbiform; two shorter ones lanceolate; ovary asymmetric, elliptic, style very short, stigma pilose, white when fresh, brown when dried. Staminate flowers inflorescent in Spring, pistillate flowers inflorescent after staminate ones. Achenes flat, ovate, pale yellow, spotted white, covered with tick gel when watered for the first time.

Distribution. Warm temperate to subtropical regions in East Asia: Japan, South Korea, Taiwan, and Mainland China.

Habitat. Bright forest edges or gaps in forests. Considering old specimens, previously common near to villages about one century ago, but currently rare around villages.

Note. A specimen (China. Hubei: T'cen-Ju-Ho, 20 Apr. – 1 May 1906, *Silvestri C. 414*, FI 018022 digital image!) is designated as a kind of its types by someone, but this is not correct, because the illustration in the original description is considered as the holotype.

Representative specimens examined. **Japan**. Kochi: Kami, Monobe, 17 Apr. 2018, *S. Aoki 713* (TI). **Mainland China**. Zhejiang: Chekiang, 20 Apr. 1934, *S. Chen 2857* (TI). **South Korea**. South Chungcheong: Taean, Baekhwasan, 4 May 1913, *T. Nakai s.n* (TI). **Taiwan**. New-Taipei: Bali, 23 Feb. 2018, *S. Aoki 604* (TI).

2. *Nanocnide zhejiangensis* X.F. Jin & Y.F. Lu, Nor. J. Bot. 37. 2019. *Typus.* China. Zhejiang: Mt. Tongling, Meichang, 16 Apr. 2012, *X. F. Jin 2806* (holo- HTC; iso- PE, ZM).

Description. Stems glabrous, erect. Leaves opposite, deltate at higher nodes, pubescent, dentate, with 11–14 teeth on edge of largest leaf; petioles elongate. Stipules opposite. Flowers diclinous. Staminate inflorescence determinate, umbel or dichasium; peduncles elongate above highest leaves; perianths glabrous, five, closed, forming turbinate buds when immature, open when matured. Pistillate inflorescence determinate umbel, solitary; perianths four.

Distribution. Endemic to Mainland China (Mt. Tongling in Zhejiang).

3. ***Nanocnide lobata*** Weddell, Prodr. 16(1): 69. 1869. *Typus.* Insulis sinensibus Loo-Choo, *C. Wright s.n.* (potential holo- US 00090518 digital image!) See also *Note* below.

Description. Annual herbs. Stems with depressed hairs. Leaves opposite when young, alternate when matured, flabellate to ovate, pubescent, with 5 teeth on edge for the largest leaf; petioles elongate. Stipules opposite, persistent after leaf deciduation. Flowers diclinous. Staminate inflorescence determinate; perianths pubescent, closed, forming turbinate buds when immature, open when matured; filaments white, translucent; pistillode circularly hollowed at center. Pistillate inflorescence determinate umbel, solitary on axils, monochasium on apices; perianths four, unequal in size: two longer ones lance-ovate, cymbiform; two shorter ones lanceolate; style very short; ovary asymmetric, elliptic; stigma pilose, white when fresh, brown when dried. Staminate flowers inflorescent in early Achenes flat, ovate, pale yellow, spotted white, covered with gel when watered for the first time.

Distribution. Vietnam, Mainland China, Ryukyu Islands in Japan, and Kyushu in Japan.

Note. Weddell (1869) cited a specimen without a collector number in the original description (Loo-Choo, *Wright s.n.*). On the other hand, he cited a specimen with a number in the description of *N. japonica* (Loo-Choo, *Wright 301*) (Weddell 1869), and I found that this specimen corresponded to *N. lobata*. If I assume that Weddell cited the specimen (Loo-Choo, *Wright 301*) by mistake in the description of *N. japonica* and that he omitted to add the collector number in the description of *N. lobata*, all the specimens (Loo-Choo, *Wright 301*; Loo-Choo, *Wright s.n.*) are syntypes. The possible syntypes except the potential holotype are listed as follows: Ins. Sinensibus Loo-Choo, *C. Wright 301* (P 06456110 digital image!). Loo-Choo, C. *Wright 301* (GH 00589550 digital image!). Loo-Choo, *C. Wright 301* (K 000708596 digital image!).

Representative specimens examined. **Japan**. Kagoshima: Okinoerabu Isl., Wadomari, 14 Mar. 2018, *S. Aoki 622* (TI). Okinawa: Ishigaki Isl., Fukai, 12 Mar. 2018, *S. Aoki 609* (TI). Yonaguni Isl., Kubura, 19 Jap. 1973, *Miyoshi Furuse 2091* (TAI 203357).

3a. subsp. ***lobata*** –Fig. 4.9. and Fig. 4.12.

Description. Stems, decumbent or ascending. Leaves, mostly dentate, entire at some lower nodes and some highest nodes, with 5 teeth on edge for the largest leaf. Stipules smaller than 0.1 cm long. Staminate inflorescence determinate, almost sessile; perianths

pubescent, four, green or white translucent; anthers white, translucent; pistillode white translucent, quadrangular. Pistillate inflorescence determinate umbel, solitary on axils, monochasium on apices; peduncles and pedicels too short to observe without dissection; perianths densely hirsute, elongate and terete in fruition. Staminate flowers inflorescent in early Spring, pistillate flowers and a small number of staminate flowers inflorescent successively.

Distribution. Ryukyu Islands in Japan.

3b. subsp. ***pilosa*** (Migo) S. Aoki & M. Ito, **comb. & stat. nov.**

Basionym: *N. pilosa* Migo, Trans. Nat. Hist. Soc. Taiwan 24: 386. 1934. *Typus*. Kunshan, 10 June 1934, *H. Migo s.n.* (holo- TI!). –Fig. 4.10. and Fig. 4.13.

Description. Stems, ascending to erect. Leaves, dentate, with 11–15 teeth on edge of largest leaf; petioles grooved on adaxial surfaces. Stipules, ovate to lanceolate, ca. 1 mm long. Staminate inflorescence determinate umbel, on upper axils; peduncles elongate at level of highest leaves for those before pistillate flowers inflorescent; pedicels ca. 1 mm long; bracts membranous, lanceolate; perianths strigose, five rarely four, upper surfaces green; anthers white; pistillode pentagonal. Pistillate inflorescence determinate umbel, solitary on axils, monochasium on apices; peduncles almost sessile, rarely elongate; pedicels too short to observe without dissection; perianths having long spikes on midribs. Staminate flowers inflorescent in Spring, pistillate flowers and a small number of staminate flowers inflorescent successively.

Distribution. Japan (only in Kagoshima Prefecture), Mainland China, and Vietnam.

Habitat. Humid forest sides.

4.6 Figures and Tables

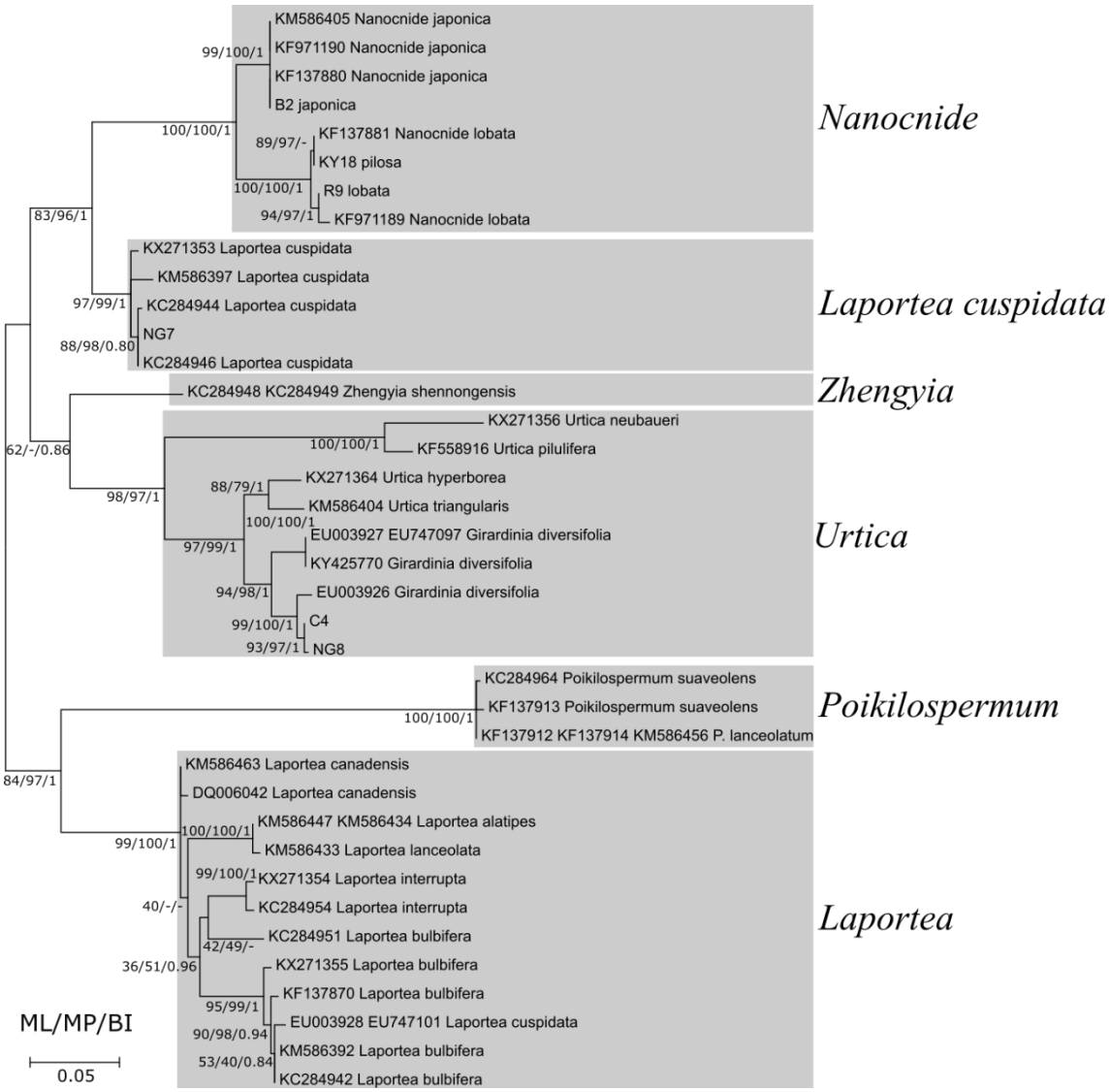

Figure 4.1. The ML tree of ITS for *Nanocnide* and its related genera. Numbers near a branch stand for ML bootstrap value, MP bootstrap value, and BI posterior probability of the branch. Hyphens mean the branch did not appear in the corresponding analysis. Samples with identical sequences are collapsed and listed in the same line except those in *Nanocnide* clade. Collapsed samples in *Nanocnide* clade are the same as Fig. 2, and can be checked in it.

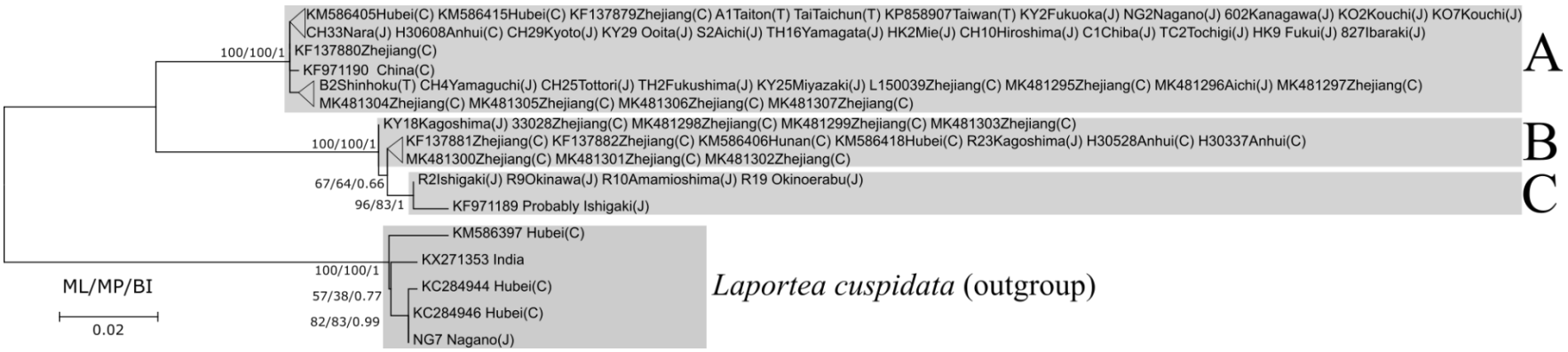

Figure 4.2. The ML tree of ITS sequence for *Nanocnide*. Numbers near a branch stand for ML bootstrap value, MP bootstrap value, and BI posterior probability of the branch. Hyphens mean the branch did not appear in the corresponding analysis. Samples with identical sequences are collapsed and listed in the same line, but a triangle on a tip of the tree shows a single sequence corresponding to the samples on listed in two lines. Abbreviation after the sample locality shows the country of the locality. (C): Mainland China, (J): Japan, and (T): Taiwan.

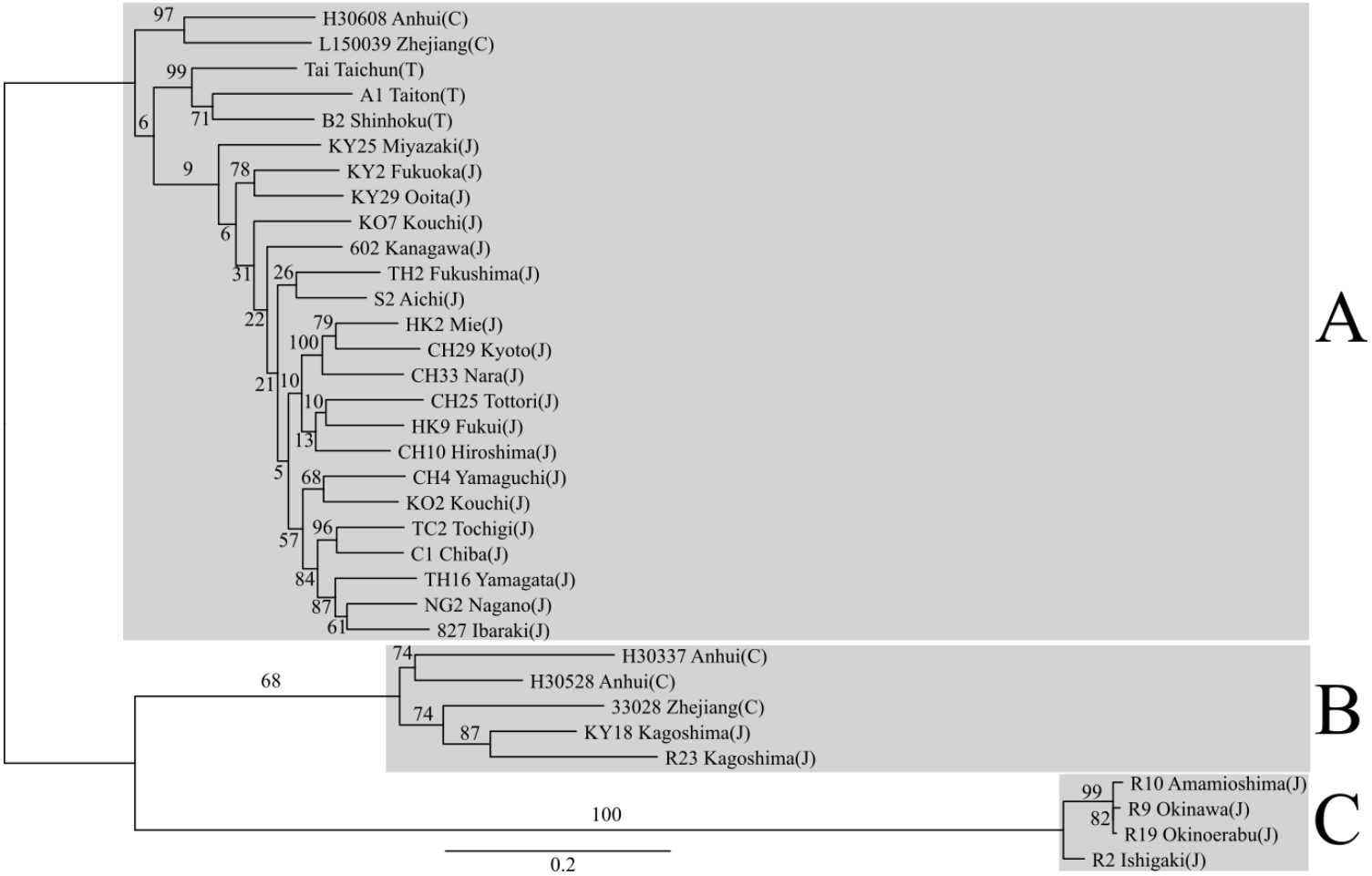

Figure 4.3. The ML tree based on MIG-seq. Numbers near a branch stand for bootstrap values (1000 times). Abbreviation after the sample locality shows the country of the locality. (C): Mainland China, (J): Japan, and (T): Taiwan.

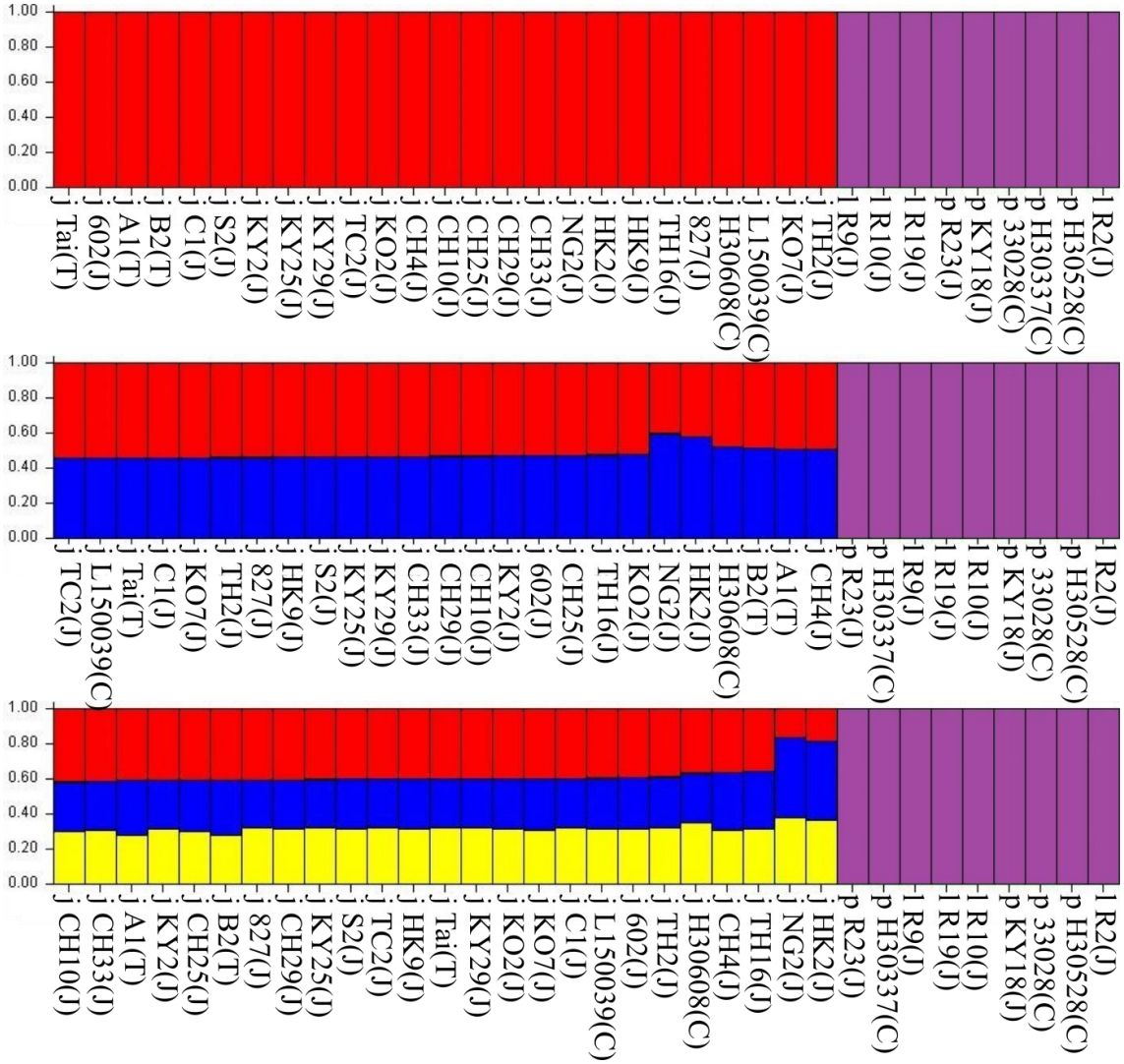

Fig.4.4. Genetic structures in *Nanocnide* inferred using Structure for r = 0.9. The top, middle and bottom graphs correspond to k = 2, 3 and 4. The alphabets j, l and p before sample names correspond to taxa of the samples: *N. japonica*, *N. lobata* and *N. pilosa*, respectively.

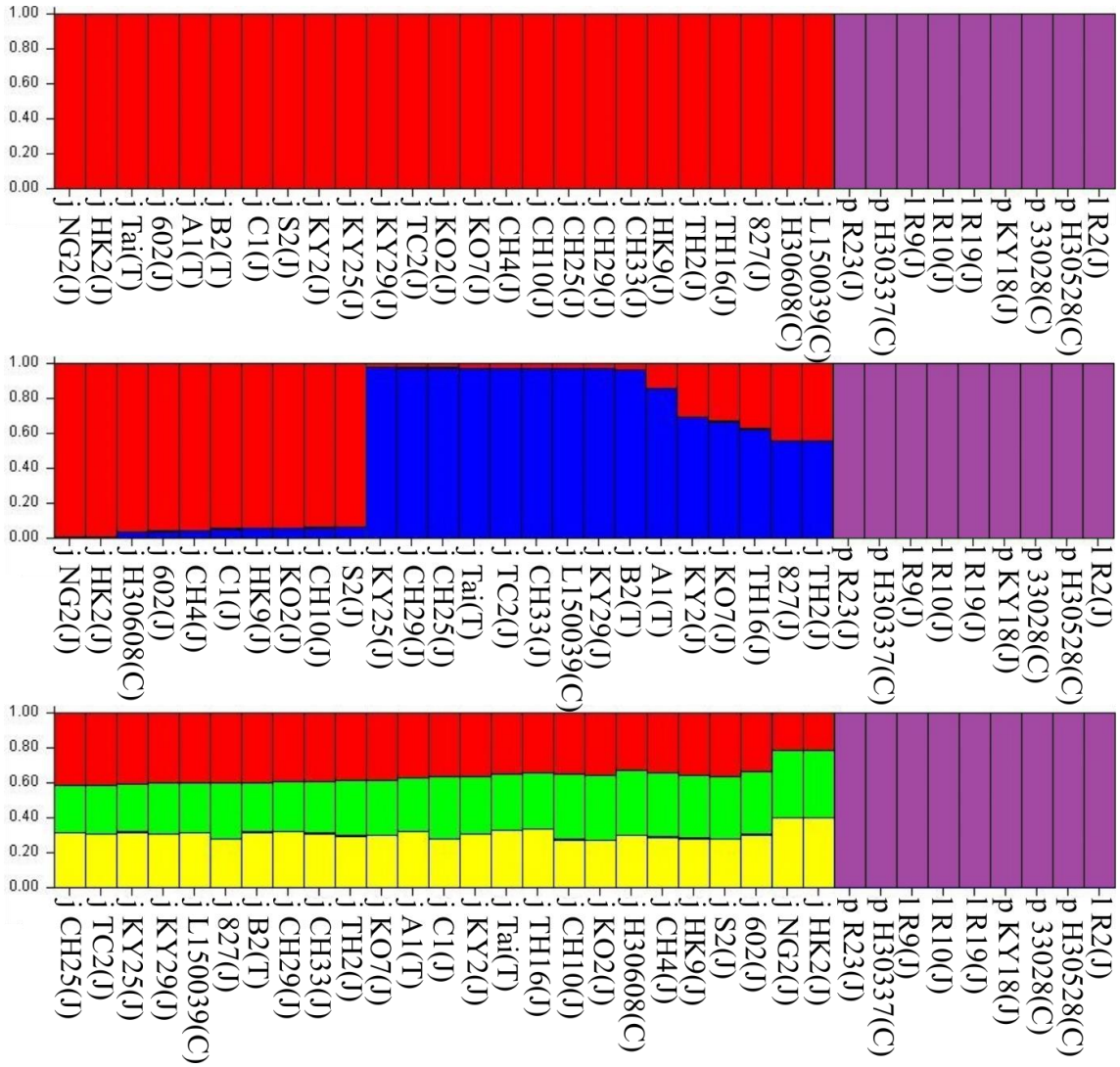

Fig.4.5. Genetic structures in *Nanocnide* inferred using Structure for r = 0.8. The top, middle and bottom graphs correspond to k = 2, 3 and 4. The alphabets j, l and p before sample names correspond to taxa of the samples: *N. japonica*, *N. lobata* and *N. pilosa*, respectively.

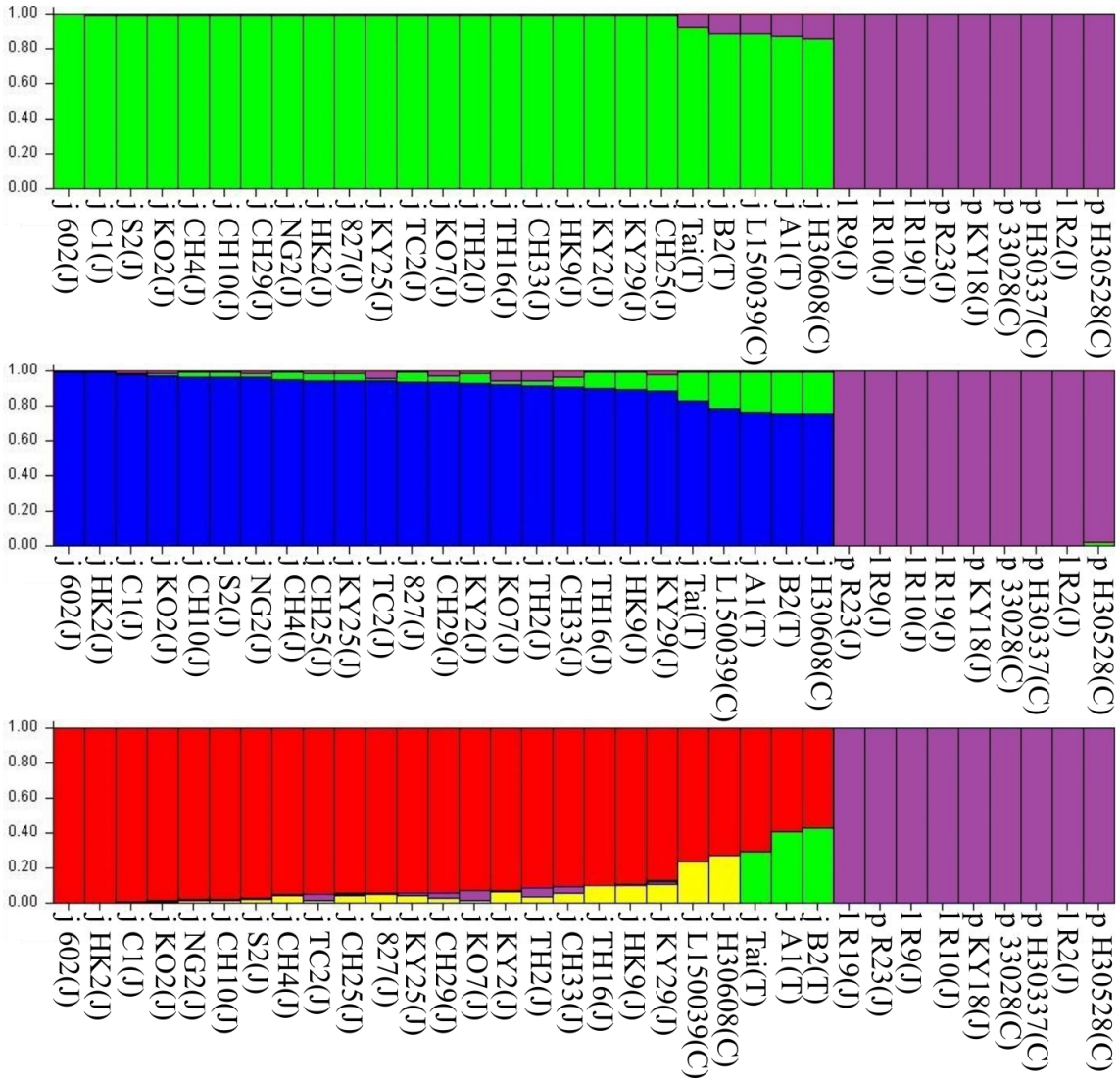

Fig.4.6. Genetic structures in *Nanocnide* inferred using Structure for r = 0.7. The top, middle and bottom graphs correspond to k = 2, 3 and 4. The alphabets j, l and p before sample names correspond to taxa of the samples: *N. japonica*, *N. lobata* and *N. pilosa*, respectively.

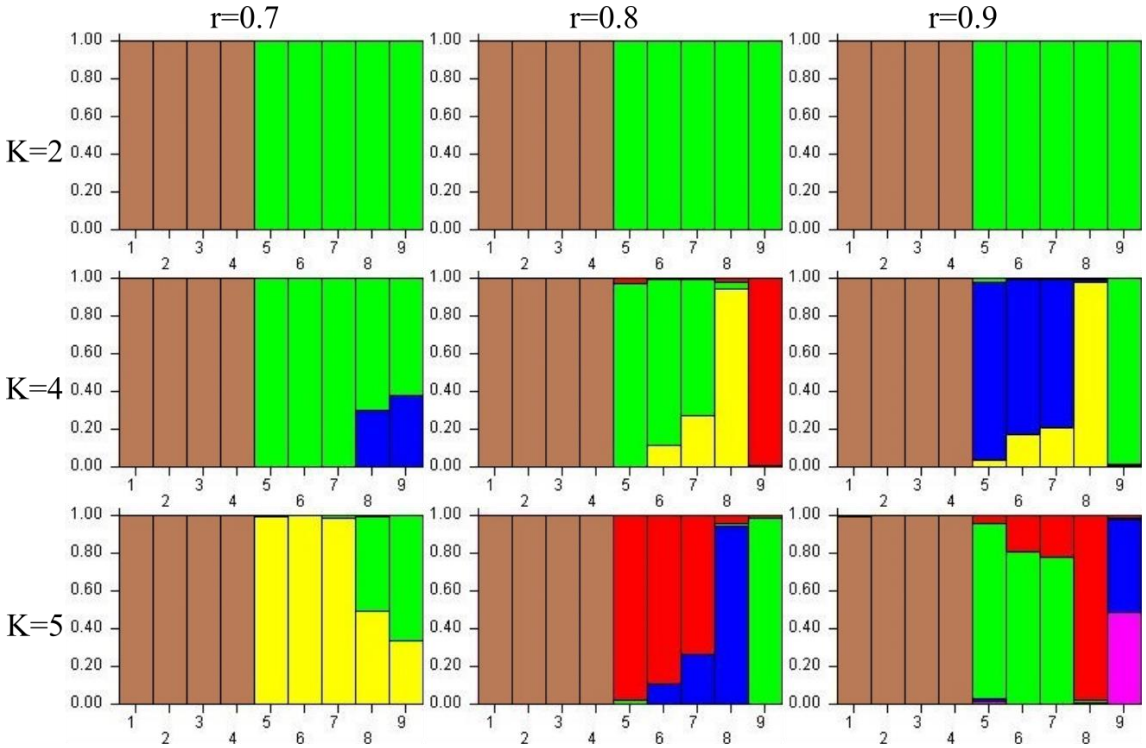

Fig.4.7. Genetic structures in *N. lobata* and *N. pilosa* inferred using Structure. The numbers on the X axes correspond to the following samples. 1–4: *N. lobata*. 5–9: *N. pilosa*. 1: R2 (Japan); 2: R9 (Japan); 3: R10 (Japan); 4: R19 (Japan); 5: R23 (Japan). 6: KY18 (Japan); 7: 33028 (Mainland China); 8: H30528 (Mainland China). 9 H30337 (Mainland China).

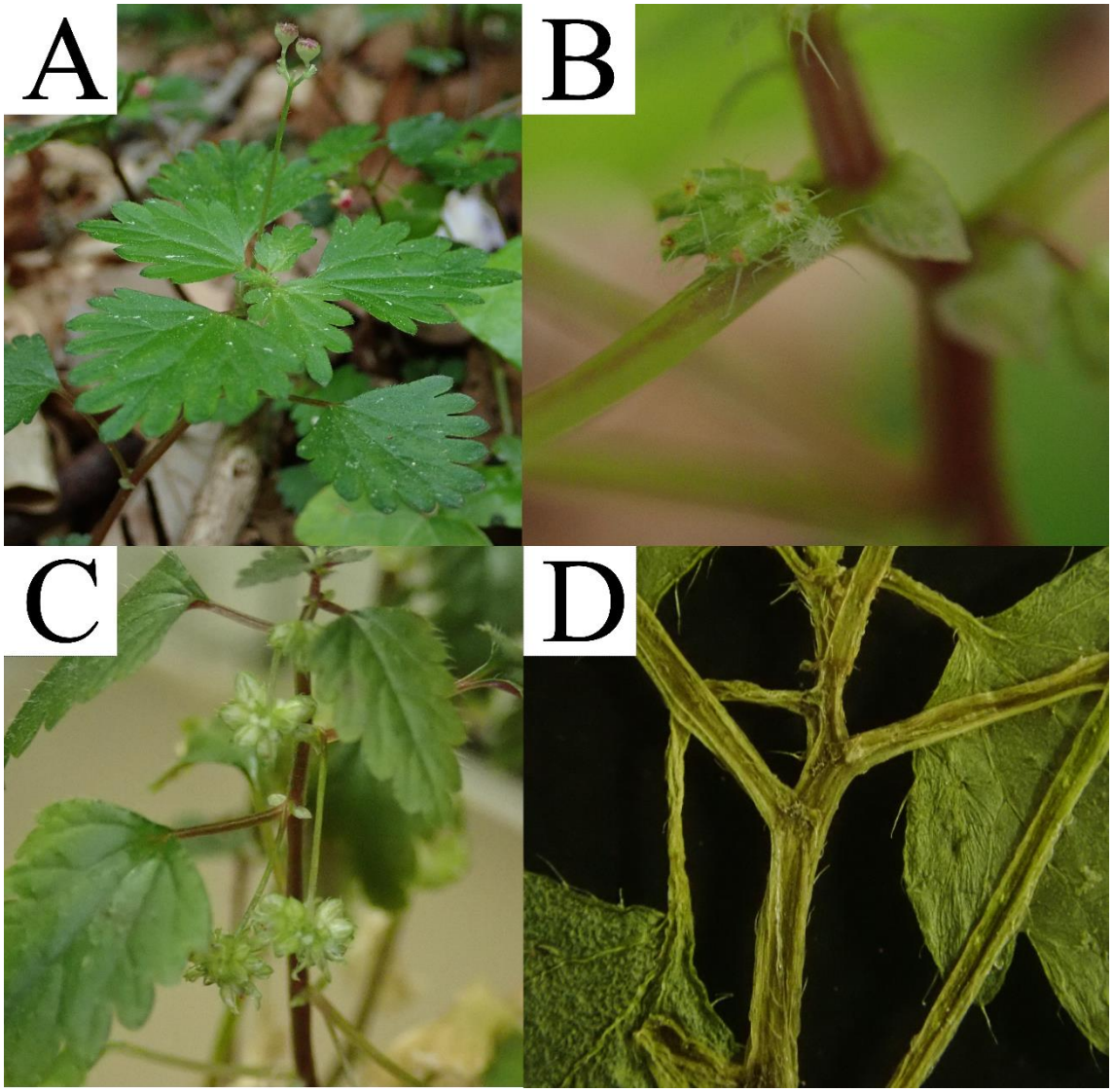

Fig. 4.8. *Nanocnide japonica*. A: Habitat and male inflorescence. B: Young female inflorescence. C: Aged female inflorescence. D: Stem hairs. (A: Japan, Fukushima Pref., Mt. Futatsutuya, 11 May 2018. B: Japan, Kochi Pref., Kami, 17 April 2018. C: In cultivation, originally collected on Taiwan, Taipei City, Bali. D. Specimen from Taiwan, Taipei City, Bali, 23 February 2018.)

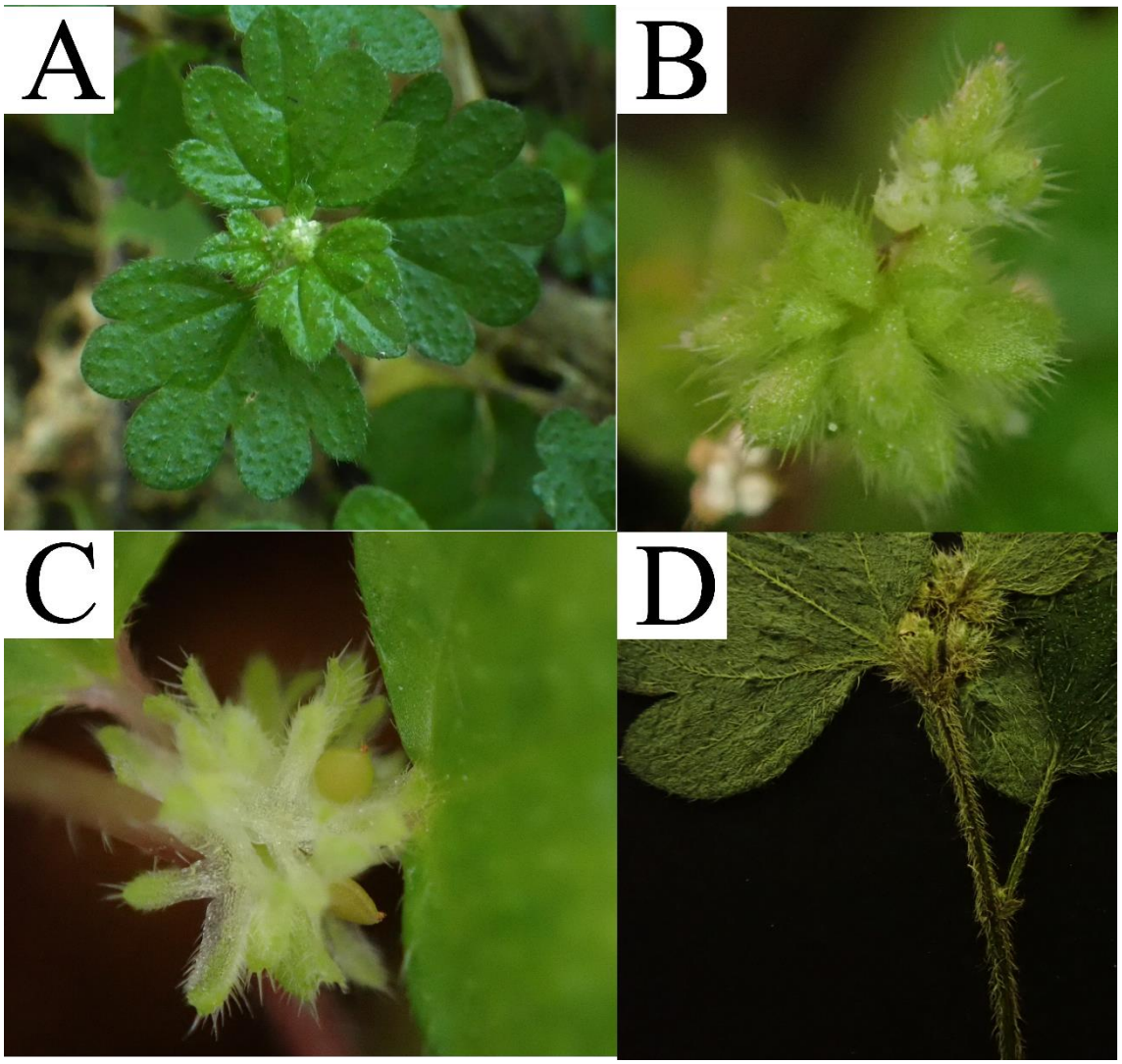

Fig. 4.9. *Nanocnide lobata* subsp. *lobata*. A: Habitat and male inflorescence. B: Young female inflorescence. C: Aged female inflorescence and capsules. D: Stem hairs. (A: Japan, Okinawa Pref., Ishigaki Isl., 12 March 2018. B–C: In cultivation, originally collected on Japan, Okinawa Pref., Okinawa Isl. D. Specimen from Japan, Okinawa Pref., Ishigaki Isl., 12 March 2018.)

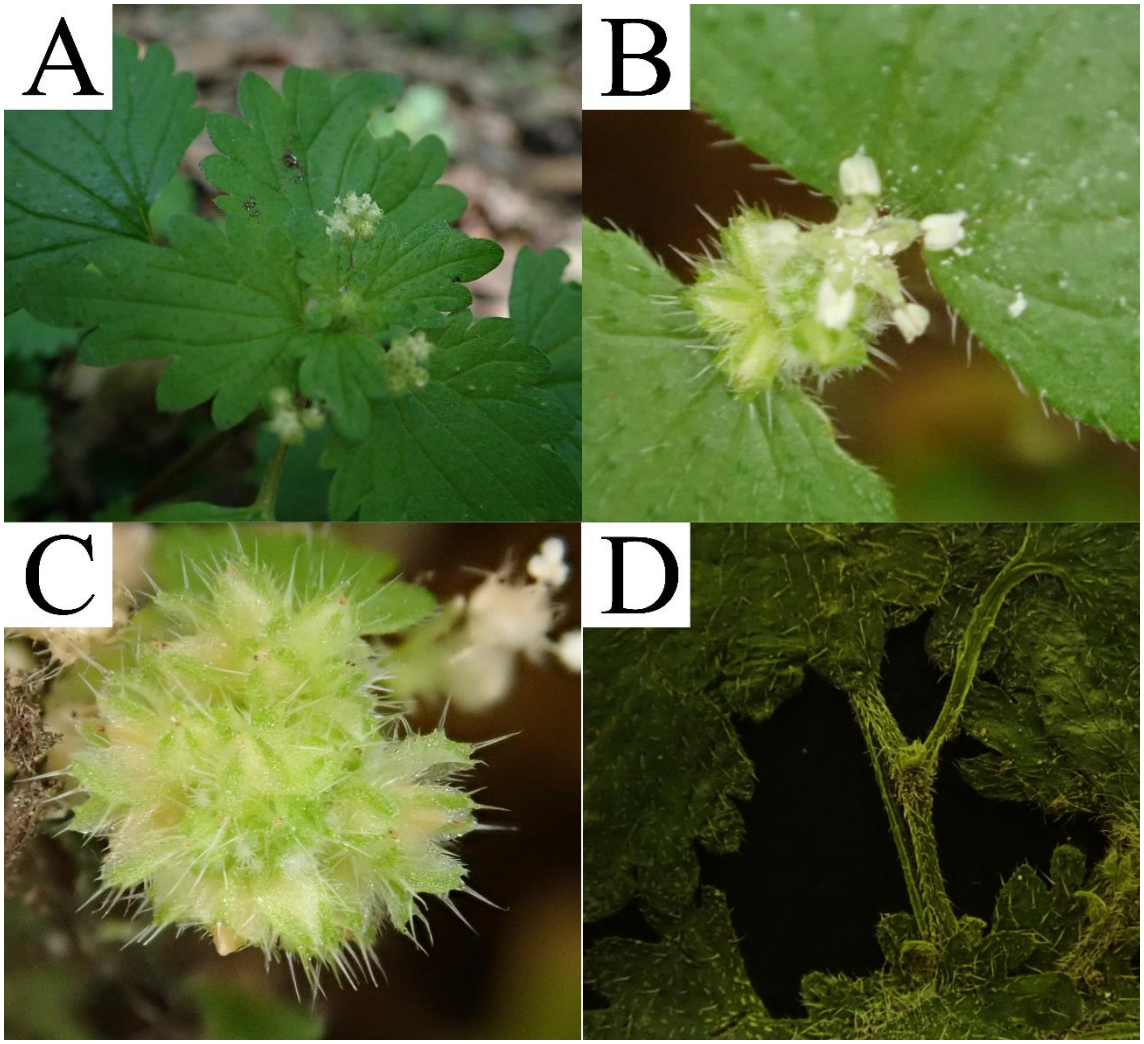

Fig. 4.10. *Nanocnide lobata* subsp. *pilosa*. A: Habitat and male inflorescence. B: Young mixed inflorescence. C: Aged female inflorescence and capsules. D: Stem hairs. (A: Japan, Kagoshima Pref., Gokabeppu, 10 April 2018. B–C: In cultivation, originally collected on Japan, Kagoshima Pref., Gokabeppu. D. Specimen from Japan, Kagoshima Pref., Ishiki, 15 March 2018.)

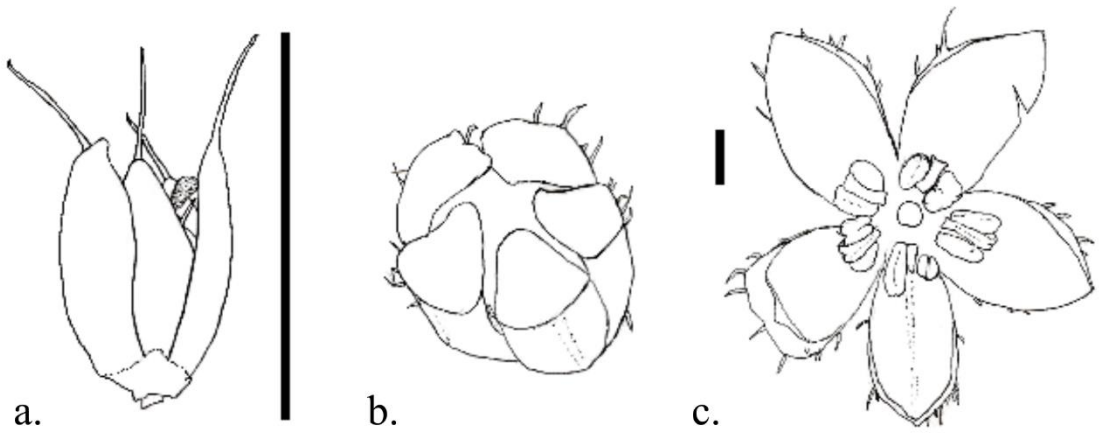

Fig. 4.11. Analysis of flowers of *N. japonica*. a. Female flower; b. Outside of male flower; c. Inside of male flower. Scale bars in a. and b. show ca. 2 mm and 1mm, respectively, and b. is drawn at almost the same scale as c. All drawings are from freeze dried specimens. (a. Japan, Tochigi Pref., Mt. Futamata. b. and c. Japan, Chiba Pref., Katori.)

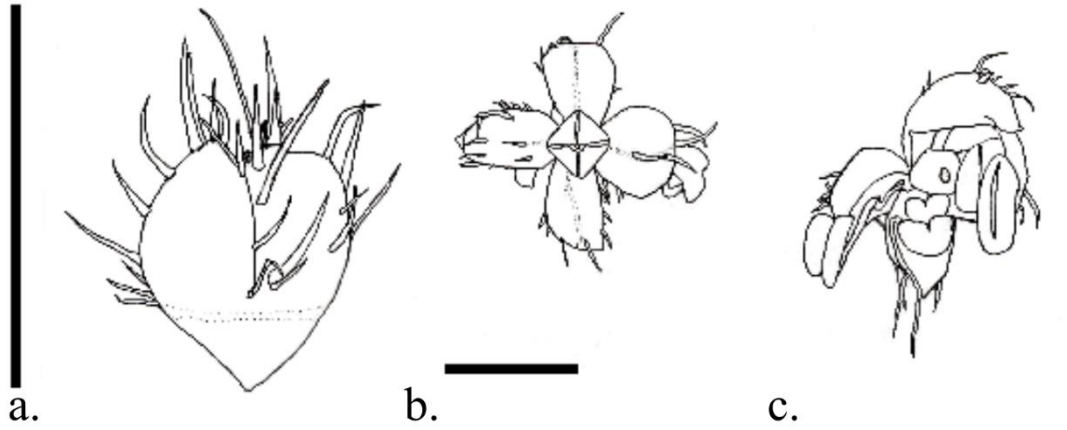

Fig. 4.12. Analysis of flowers of *N. lobata* subsp. *lobata*. a. Female flower; b. Outside of male flower (observed from basal direction); c. Inside of male flower. Both of the scale bars in a. and b. show ca. 1 mm, and c. is drawn at almost the same scale as b. All drawings are from freeze dried specimens. (a. Japan, Kagoshima Pref., Okinoerabu Isl. b. and c. Japan, Okinawa Pref., Okinawa Isl.)

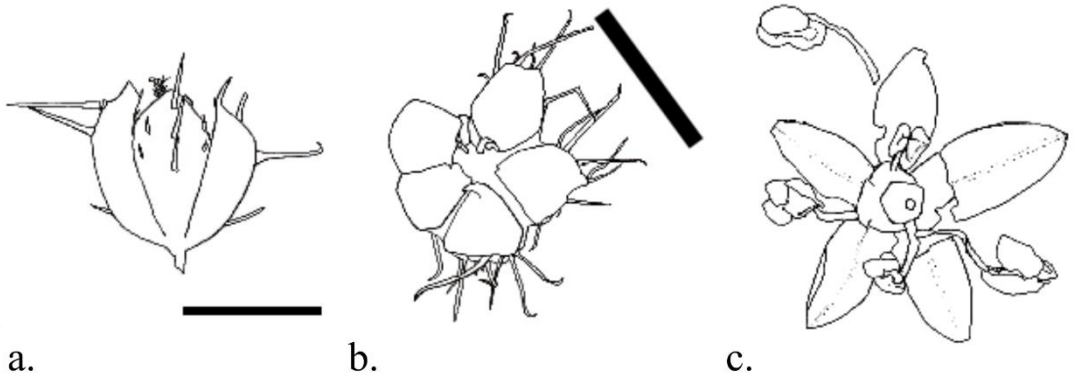

Fig. 4.13. Analysis of flowers of *N. lobata* subsp. *pilosa*. a. Female flower; b. Outside of male flower; c. Inside of male flower. Both of the scale bars in a. and b. show ca. 1 mm, and c. is drawn at almost the same scale as b. Note that hairs on left two perianths in b. are omitted. All drawings are from freeze dried specimens. (a., b. and c. Japan, Kagoshima Pref., Ishiki.)

Table 4.1. Existing classification of *Nanocnide*.

Basionym	Jiarui et al. (2003)	Tateishi (2006)
N. japonica	*N. japonica*	*N. japonica*
N. dichotoma	*= N. japonica*	*= N. japonica*
N. lobata	*N. lobata*	*N. lobata*
N. pilosa	*= N. lobata*	*N. pilosa*
N. zhejiangensis	Not treated	Not treated

Table 4.2. Analyzed materials. Abbreviations in Locality: C= Mainland China; J= Japan; T= Taiwan.

Taxon	Abbr. in text	Locality	Specimen Information	Accession No.
N. lobata	R2	Ishigaki Isl., Okinawa, J.	S. Aoki 608; TI	LC507645
	R9	Okinawa Isl., Okinawa, J.	S. Aoki 615; TI	LC507647
	R10	Amami-oshima, Kagoshima, J.	S. Aoki 616; TI	LC507643
	R19	Okinoerabu Isl., Kagoshima, J.	S. Aoki 625; TI	LC507644
N. pilosa	33028	Yuyao, Zhejiang, C.	Yuyao exploration team; HZU	LC507616
	H30528	Huangshan, Anhui, C.	Li et al.; HZU	LC507629
	H30337	Huangshan, Anhui, C.	Li et al.; HZU	LC507628
	R23	Ishiki, Kagoshima, J.	S. Aoki 629; TI	LC507646
	KY18	Gokabeppu, Kagoshima, J.	S. Aoki 669; TI	LC507635
N. japonica	Tai	Heping, Taichun, T.	S. Aoki 480; TI	LC507649
	602	Hadano, Kanagawa, J.	S. Aoki 602; TI	LC507617
	A1	Haiduan, Taitung, T.	S. Aoki 603; TI	LC507619
	B2	Bali, New Taipei, T.	S. Aoki 605; TI	LC507620
	C1	Katori, Chiba, J.	S. Aoki 636; TI	LC507621
	S2	Toyohashi, Aichi, J.	S. Aoki 647; TI	LC507648
	KY2	Mt. Rai, Fukuoka, J.	S. Aoki 653; TI	LC507636
	KY25	Miyazaki, Miyazaki, J.	S. Aoki 676; TI	LC507637
	KY29	Megusuno, Ohita, J.	S. Aoki 680; TI	LC507638
	TC2	Mt. Futamata, Tochigi, J.	S. Aoki 706; TI	LC507650
	KO2	Kami, Kochi, J.	S. Aoki 711; TI	LC507633
	KO7	Shimanto, Kochi, J.	S. Aoki 716; TI	LC507634
	CH4	Sozukyo, Yamaguchi,	S. Aoki 737; TI	LC507627

		J.		
	CH10	Shobara, Hiroshima, J.	S. Aoki 743; TI	LC507623
	CH25	Chizu, Tottori, J.	S. Aoki 758; TI	LC507624
	CH29	Kyoto, Kyoto, J.	S. Aoki 762; TI	LC507625
	CH33	Kawakami, Nara, J.	S. Aoki 766; TI	LC507626
	NG2	Komoro, Nagano, J.	S. Aoki 770; TI	LC507640
	HK2	Inabe, Mie, J.	S. Aoki 778; TI	LC507631
	HK9	Mt. Monju, Fukui, J.	S. Aoki 785; TI	LC507632
	TH2	Mt. Futatsuya, Fukushima, J.	S. Aoki 809; TI	LC507652
	TH16	Yamadera, Yamagata, J.	S. Aoki 823; TI	LC507651
	827	Mt. Tsukuba, Ibaraki, J.	S. Aoki 827; TI	LC507618
	H30608	She, Anhui, C.	Li et al.; HZU	LC507630
	L150039	Lin'an, Zhejiang, C.	P. Li 150039; ZHU	LC507639
(outgroup)				
L. cuspidata	NG7	Komoro, Nagano, J.	S. Aoki 775; TI	LC507641
U. thunbergiana	C4	Katori, Chiba, J.	S. Aoki 639; TI	LC507622
	NG8	Komoro, Nagano, J.	S. Aoki 776; TI	LC507642

Table 4.3. Sequences in GenBank used in this study. Study and Locality of the outgroup were omitted. The classification of outgroup followed the registered data, even when they were dubious or contained synonyms.

Taxon	Study	Locality	Accession No.
N. japonica	Henning et al. (2014)	China	KF971190
	Wu et al. (2003)	Zhejiang, China	KF137879
	Wu et al. (2003)	Zhejiang, China	KF137880
	Kim et al. (2016)	Hubei, China	KM586405
	Kim et al. (2016)	Hubei, China	KM586415
	Tseng et al. (2019)	Taiwan	KP858907
	Jin et al. (2019)	Zhejiang, China	MK481295

	Jin et al. (2019)	Aichi, Japan	MK481296
	Jin et al. (2019)	Zhejiang, Japan	MK481297
N. lobata	Henning et al. (2014)	Japan	KF971189
N. pilosa	Wu et al. (2003)	Zhejiang, China	KF137881
	Wu et al. (2003)	Zhejiang, China	KF137882
	Kim et al. (2016)	Hunan, China	KM586406
	Kim et al. (2016)	Hubei, China	KM586418
	Jin et al. (2019)	Zhejiang, China	MK481298
	Jin et al. (2019)	Zhejiang, China	MK481299
	Jin et al. (2019)	Zhejiang, China	MK481300
	Jin et al. (2019)	Zhejiang, China	MK481301
	Jin et al. (2019)	Zhejiang, China	MK481302
	Jin et al. (2019)	Zhejiang, China	MK481303
N. zhejiangensis	Jin et al. (2019)	Zhejiang, China	MK481304
	Jin et al. (2019)	Zhejiang, China	MK481305
	Jin et al. (2019)	Zhejiang, China	MK481306
	Jin et al. (2019)	Zhejiang, China	MK481307
(outgroup)			
L. cuspidata	-	-	KX271353
	-	-	KM586397
	-	-	KC284944
	-	-	EU003928
	-	-	KC284946
Zhengyia shennongensis	-	-	KC284948
	-	-	KC284949
Urtica neubaueri	-	-	KX271356
U. pilulifera	-	-	KF558916
U. hyperborea	-	-	KX271364
U. triangularis	-	-	KM586404
Girardinia diversifolia	-	-	EU003927
	-	-	KY425770
	-	-	EU003926
Poikilospermum suaveolens	-	-	KC284964

	-	-	KF137913
	-	-	KF137914
	-	-	KM586456
P. lanceolatum	-	-	KF137912
L. canadensis	-	-	KM586463
	-	-	DQ006042
L. alatipes	-	-	KM586447
	-	-	KM586434
L. lanceolata	-	-	KM586433
L. interrupta	-	-	KX271354
	-	-	KC284954
L. bulbifera	-	-	KC284951
	-	-	KF137870
	-	-	KM586392
	-	-	KC284942
L. macrostachya	-	-	EU747101
	-	-	EU747097

Table 4.4. Natural logarithm of the mean of estimated likelihood for each r and k in the structure analyses using all the *Nanocnide* samples.

		k				
		1	2	3	4	5
r	0.9	-323.7	-67.3	-62.8	-65.0	-69.8
	0.8	-410.5	-113.6	-87.9	-90.3	-97.4
	0.7	-2414.0	-1825.2	-1858.9	-1803.3	-1951.4

Table 4.5. Natural logarithm of the mean of estimated likelihood for each r and k in the structure analyses using *N. lobata* and *N. pilosa* samples.

		k				
		1	2	3	4	5
r	0.9	-1966.9	-60.5.5	-492.0	-393.8	-402.0
	0.8	-3689.6	-1149.4	-1149.0	-811.8	-821.1
	0.7	-5641.6	-1860.9	-12182.3	-13183.0	-1690.5

Chapter 5. Concluding remarks

This study suggested a remedy for difficulty in sampling of wild taxa. The presented solution contains theorization of the conventional sampling method by taxonomists and departure from frequentistic sampling theory. While the test of the new sampling theory, spatial sampling, against existing data suggested its wide applicability in spite of its theoretical simplicity, its practical application to the present taxonomic study to *Nanocnide* demonstrated many points to improve for spatial sampling. Representative statistics of characteristics of wild taxa which are based on spatial sampling are also still undefined. I devoutly hope that many and various researchers step into this frontier and develop theoretical foundation of filed sciences treating wild taxa to increase their accuracy, objectivity and reproducibility.

Acknowledgements

First of all, I would like to thank Prof. Motomi Ito for his suggestions in researches and generosity to my research which goes out of the realm of plant taxonomy.

For the research in Chapter 2, I thank Dr. Kenji Tani in Saitama University for his website (http://ktgis.net/gcode/lonlatmapping.html) which was used for debugging Samploc. For the research in Chapter 3, I thank Dr. Fumio Tajima for his advice on the English expressions and Dr. Masakazu Shimada for highlighting a miscalculation in the manuscript. For the research in Chapter 4, I would like to thank curators of herbaria MAK, MBK, TAI and TI for allowing me to observe specimens in their care. I also thank Mr. Chen-Jui Yang and Mr. Yu-Chun Liu at National Taiwan University for their help in sampling in Taiwan.

I also thank all colleagues in Ito laboratory for their numerous advices and aids.

The researches in Chapter 2 and 4 were supported by the Sasakawa Scientific Research Grant from the Japanese Science Society (2018-5044). The research in Chapter 3 was supported by AMED (16km0210053j0005).

References

Algina J., Keselman H. & Penfield R. (2005) Alternative to Cohen's standardized mean difference effect size: robust parameter and confidence interval in the two independent group case. *Psychological Methods* 10 (3): 317–328.

American Educational Research Association. (2006) Standards for reporting on empirical social science research in AERA publication. *Educational Researcher*, 35 (6): 33–40.

American Psychological Association. (2009) *Publication manual of the American Psychological Association* (6th ed.). American Psychological Association.

Barbujani G. (1987) Autocorrelation of gene frequencies under isolation by distance. *Genetics* 117 (4): 777–782.

Bergten J. (2005) A review of long-branch attraction. *Cladistics* 21 (2):163–193.

Blume C. L. von. (1856) *Museum Botanicum Lugduno-Batavum, sive, stirpium exoticarum novarum vel minus cognitarum ex vivis aut siccis brevis ex positio*, 2. E. J. Brill, Lugdunum-Batavorum. 256 pp.

Bohonak A. J. (2002) IBD (isolation by distance): A program for analyses of isolation by distance. *Journal of Heredity* 93 (2): 153–154.

Bonett D. G. (2008) Confidence intervals for standardized linear contrasts of means. *Psychological Methods* 13 (2): 99–109.

Borchsenius F. (2009) FastGap homepage, FastGap 1.2., Department of Biosciences, Aarhus University, Denmark. Available from: http://www.aubot.dk/FastGap_home.htm (accessed: 12 July 2019).

Catchen J., Hohenlohe P. A., Bassham S., Amores A. & Cresko W. A. (2013) Stacks: an analysis tool set for population genomics. *Molecular Ecology* 22 (11): 3124–3140.

Chebyshev P. (1867) Des Vaeurs Moyennes. *Journal de Mathématiques Pures et Appliquées* 12: 177–184.

Chessel D., Dufour A. & Thioulouse J. (2004) The ade4 package - I: One-table methods. *R News* 4 (1): 5–10.

Cohen J. (1962) The statistical power of abnormal-social psychological research: A review. *The Journal of Abnormal Psychology* 65 (3): 145–153.

Cohen J. (1969) *Statistical Power Analysis for the Behavioral Sciences*. Academic Press.

Cohen J. (1988) *Statistical Power Analysis for the Behavioral Sciences second edition*. Lawrence Erlbaum Associates, USA.

Cumming G. & Finch S. (2001) A primer on the understanding, use and calculation of confidence intervals that are based on central and non-central distributions.

Educational and Psychological Measurement 61 (4): 532–574.

Deng T., Kim C., Zhang D.-G., Zhang J.-W., Li Z.-M., Nie Z.-L. & Sun H. (2013) *Zhengyia shennongensis*: A new bulbiferous genus and species of the nettle family (Urticaceae) from central China exhibiting parallel evolution of the bulbil trait. *Taxon* 62 (1): 89–99.

Der J. P., Thomson J. A., Stratford J. K. & Wolf P. G. (2009) Global chloroplast phylogeny and biogeography of bracken (*Pteridium*; Dennstaedtiaceae). *American Journal of Botany* 96 (5): 1041–1049.

Dunlap W., Cortina J., Vaslow J. & Burke M. (1996) Meta-analysis of experiments with matched groups or repeated measures designs. *Psychological Methods* 1 (2): 170–177.

Earl D. A. & vonHoldt B. M. (2012) Structure Harvester: a website and program for visualizing Structure output and implementing the Evanno method. *Conservation Genetics Resources* 4(2): 359–361.

Emerson K. J., Merz C. R., Catchen J. M., Hohenlohe P. A., Creskok W. A., Bradshaw W. E., & Holzapfel C. M. (2010) Resolving postglacial phylogeography using high-throughput sequencing. *PNAS* 107 (37): 16196–16200.

Faith D. P. & Walker P. A. (1996) Environmental diversity: on the best-possible use of surrogate data for assessing the relative biodiversity of sets of areas. *Biodiversity and Conservation* 5 (4): 399–415.

Faurby S., Jørgensen A., Kristensen R. M. & Funch P. (2011) Phylogeography of North Atlantic intertidal tardigrades: refugia, cryptic speciation and the history of the Mid-Atlantic islands. *Journal of Biogeography* 38 (8): 1613–1624.

Fisher R. (1925) *Statistical Methods for Research Workers*, Oliver and Boyd.

Fisher R. (1936) The use of multiple measurements in taxonomic problems. *The Annals of Eugenics* 7 (2): 179–188.

Forsskål P. (1775) *Flora Aegyptiaco-Arabica. sive, descriptions plantarum, quas per Aegyptum inferiorem et Arabiam felicem*. Officina Mölleri, Haunia.

Friis I. (1981) A synopsys of *Girardinia* (Urticaceae). *Kew Bulletin* 36 (1): 143–157.

Fujii N., Tomaru N., Okuyama K., Koike T., Mikami T. & Ueda K. (2002) Chloroplast DNA phylogeography of *Fagus crenata* (Fagaceae) in Japan. *Plant Systematics and Evolution* 232 (1–2): 21–33.

Fukuoka, N. & Kurosaki, N. (1995) Bulbils of *Laportea cuspidata* Forma *bulbifera* (Urticaceae). *Memoirs of the National Science Museum* 28: 87–90.

Geml J., Tulloss R. E., Laursen G. A., Sazanova N. A. & Taylor D. L. (2008) Evidence for strong inter- and intracontinental phylogeographic structure in *Amanita muscaria*,

a wind-dispersed ectomycorrhizal basidiomycete. *Molecular Phylogenetics and Evolution* 48 (2): 694–701.

Glass G. (1976) Primary, secondary, and meta-analysis of research. *Educational Researcher* 5 (10): 3–8.

Greene S. L. & Hart T. C. (1999) Implementing geographic analysis in germplasm conservation. In Greene S. L. & Suarino L. (Eds.), *Linking genetic resources and geography: emerging strategies for conserving and using crop biodiversity*. American Society of Agronomy; Crop Science Society of America, Madison. pp 25–38.

Grissom R. (2000) Heterogeneity of variance in clinical data. *Journal of Consulting and Clinical Psychology* 68 (1): 155–165.

Grissom R. & Kim J. (2001) Review of assumptions and problems in the appropriate conceptualization of effect size. *Psychological Methods* 6 (2): 135–146.

Grosse-Veldmann, B., Nürk N. M., Smissen R., Breitwieser I., Quandt D. & Weigend M. (2006) Pulling the sting out of nettle systematics – A comprehensive phylogeny of the genus *Urtica* L. (Urticaceae). *Molecular Phylogenetics and Evolution* 102: 9–19.

Handel-Mazzetti H. (1929) Anthophyta. *Symbolae Sinicae botanische ergebnisse der expediton der akademie der wissenshaften in wien nach Südwest-China* 7(1): 1–730.

Hedges L. V. (1981) Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics* 6 (2): 107–128.

Hedges L. V. & Olkin, I. (1985) *Statistical Methods for Meta-analysis*. Academic Press, Orlando.

Henning T., Quandt D., Grosse-Veldmann B., Monro A. & Weigend M. (2013) Weeding the nettles II: A delimitation of "*Urtica dioica* L." (Urticaceae) based on morphological and molecular data, including a rehabilitation of *Urtica gracilis* Ait. *Phytotaxa* 162 (2): 61–83.

Huaxing Q. & Gilbert M. G. (2008) Euphorbiaceae. 40. Acalypha. *In*: Wu. C., Peter H. R. & Hong D. (Eds.) *Flora of China*, 11. Science Press, Beijing & Missouri Botanical Garden Press, St. Louis, pp. 251–255.

Ibrahim K. M., Nichols R. A. & Hewitt G. M. (1996) Spatial patterns of genetic variation generated by different forms of dispersal during range expansion. *Heredity* 77: 282–291.

Iino S. (2007) Katei de tsukureru kinoko no furi-zudorai. *Chiba Mycological Club Bulletin* 22: 10–13. (in Japanese)

Iwasaki T., Aoki K., Seo A. & Murakami N. (2012) Comparative phylogeography of four component species of deciduous broad-leaved forests in Japan based on

chloroplast DNA variation. *Journal of Plant Research* 125 (2): 207–221.

Jiarui C., Friis I. & Wilmot-Dear C. M. (2003) Urticaceae. 2. Nanocnide. *In*: Wu. C., Peter H. R. & Hong D. (Eds.) *Flora of China*, 5. Science Press, Beijing & Missouri Botanical Garden Press, St. Louis, pp. 84–85.

Jiarui C. & Monro A. K. (2003) Urticaceae. 6. Pilea. *In*: Wu. C., Peter H. R. & Hong D. (Eds.) *Flora of China*, 5. Science Press, Beijing & Missouri Botanical Garden Press, St. Louis, pp. 92–120.

Jin X.-F., Zhang J., Lu Y.-F., Yang W.-W. & Chen W.-J. (2019) *Nanocnide zhejiangensis* sp. nov. (Urticaceae: Urticeae) from Zhejiang Province, East China. *Nordic Journal of Botany* 37 (10):1–7.

Johnson N. L. & Welch B. L. (1940) Applications of the non-central t-distribution. *Biometrika* 31 (3–4): 362–389.

Kageyama M., Braconnot P., Harrison S. P., Haywood A. M., Jungclaus J. H., Otto-Bliesner B. L., Peterschmitt J.-Y., Abe-Ouchi A., Albani S., Bartlein P. J., Brierley C., Crucifix M., Dolan A., Fernandez-Donado L., Fischer H., Hopcroft P. O., Ivanovic R. F., Lambert F., Lunt D. J., Mahowald N. M., Peltier W. R., Phipps S. J., Roche D. M., Schmidt G. A., Tarasov L., Valdes P. J., Zhang Q. & Zhou T. (2018). The PMIP4 contribution to CMIP6 – Part 1: Overview and over-arching analysis plan. *Geoscientific Model Development* 11: 1033–1057.

Kahan W. (1965) Further remarks on reducing truncation errors. *Communications of the ACM* 8 (1): 40.

Karney C. F. F. (2013) Algorithms for geodesics. *Journal of Geodesy* 87 (1): 43–55.

Kato Y. & Yagi T. (2004) Biogeography of the subspecies of *Parides* (*Byasa*) *alcinous* (Lepidoptera: Papilionidae) based on a phylogenetic analysis of mitochondrial ND5 sequences. *Systematic Entomology* 29 (1): 1–9.

Katoh K., Rozewicki J. & Yamada D. K. (2017) MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Briefings in Bioinformatics* bbx108.

Katoh K. & Standley M. D. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution* 30 (4): 772–780.

Kawamoto Y., Shotake T., Nozawa K, Kawamoto S, Tomari K, Kawai S, Shirai K., Morimitsu Y., Takagi N., Akaza H., Fujii H., Hagihara K., Aizawa K., Akachi S., Oi T. & Hayaishi S. (2007) Postglacial population expansion of Japanese macaques (*Macaca fuscata*) inferred from mitochondrial DNA phylogeography. *Primates* 48 (1): 27–40.

Khachaturyan A. G., Semenovskaya S. V. & Vainshtein B. K. (1979) Statistical-thermodynamic approach to determination of structure amplitude phases. *Soviet Physics Crystallography* 24: 905–916.

Kim C., Deng T., Chase M., Zhang D.G., Nie Z.-L. & Sun H. (2015) Generic phylogeny and character evolution in Urticeae (Urticaceae) inferred from nuclear and plastid DNA regions.

Kimura M. & Weiss G. H. (1964) The stepping stone model of population structure and the decrease of genetic correlation with distance. *Genetics* 49 (4): 561–576.

Kitaguchi M. (1937) On the Vegetation of Wei-hu-ling, Prov. Chi-lin, Manchuria. *Report of the Institute of Scientific Research Manchoukuo* 1(8): 255–324.

Kitamura S. & Murata G. (1962) New names and new conceptions adopted in our coloured illustrations of herbaceous plants of Japan II (Choripetalae). *Acta Phytotaxonomica et Geobotanica* 20: 195–208.

Léveillé A. A. H. (1904) Contribution jubilaire a la flore du Kouy-Tchéou. *Bulletin de la société botanique de France* 51: cxliii–cxlvi.

Linnaeus C.V. (1753) *Species Plantarum*, vol. 2. Laurentii Salvii, Holmiae.

López B. A., Tellier F., Retamal-Alarcón J. C., Pérez-Araneda K., Fierro A.O., Macaya E. C., Tala F. & Thiel M. (2017) Phylogeography of two intertidal seaweeds, *Gelidium lingulatum* and *G. rex* (Rhodophyta: Gelidiales), along the South East Pacific: patterns explained by rafting dispersal? *Marine Biology* 164: 188.

Malécot G. (1955) The decrease of relationship with distance. *Cold Spring Harbor Symposia on Quantitative Biology* 20: 52–53.

Maltagliati F., Giuseppe G. D., Barbieri M., Castlelli A. & Dini F. (2010) Phylogeography and genetic structure of the edible sea urchin *Paracentrotus lividus* (Echinodermata: Echinoidea) inferred from the mitochondrial cytochrome b gene. *Biological Journal of the Linnean Society* 100 (4): 910–923.

Mantel N. (1967) The detection of disease clustering and a generalized regression approach. *Cancer Research* 27 (2): 209–220.

Marfo P. & Okyere G. (2019) The accuracy of effect-size estimates under normals and contaminated normals in meta-analysis. *Heliyon* 5 (6): e01838.

Marsaglia G. (2003) Xorshift RNGs. *Journal of Statistical Software*, 8.

Matsumoto M. & Nishimura T. (1998) Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Transactions on Modeling and Computer Simulation* 8 (1): 3–30.

Maximowicz C. J. (1876) Diagnoses plantarum Novarum Japoniae et Mandhuriae. *Mélanges biologiques tirés du bulletin de L'académie impériale des sciences de st.*

Pétersbourg 9: 581–660.

Melnyk T. W., Knop O. & Smith W. R. (1977) External arrangements of points and unit charges on a sphere: equilibrium configurations revisited. *Canadian Journal of Chemistry* 55 (10): 1745–1761.

Migo H. (1934) Duae novae plantae Chinenses. *Transactions of the Natural History Society of Formosanum* 24: 386–388.

Minamiya Y., Yokoyama J. & Fukuda T. (2009) A phylogeographic study of the Japanese earthworm, *Metaphire sieboldi* (Horst, 1883) (Oligochaeta: Megascolecidae): Inferences from mitochondrial DNA sequences. *European Journal of Soil Biology* 45 (5–6): 423–430.

Moe A. M. & Weiblen G. D. (2012) Pollinator-mediated reproductive isolation among dioecious fig species (*Ficus*, Moraceae). *Evolution* 66 (12): 3710–3721.

Nakagawa S. & Cuthill I. (2007) Effect size, confidence interval and statistical significance: A practical guide for biologists. *Biological Reviews* 82 (4): 591–605.

Narendran T. C. (2006) *An introduction to Taxonomy*. The Director, Zoological Survey of India, Kolkata.

Nei M. & Li W. H. (1979) Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences of the United States of America* 76 (10): 5269–5273.

Nei M. & Tajima F. (1981) DNA polymorphism detectable by restriction endonucleases. *Genetics* 97(1): 145–163.

Nicholls J. A. & Austin J. J. (2005) Phylogeography of an east Australian wet-forest bird, the satin bowerbird (*Ptilonorhynchus violaceus*), derived from mtDNA, and its relationship to morphology. *Molecular Ecology* 14 (5): 1485–1496.

Nichols R. A. & Hewitt G. M. (1994) The genetic consequences of long distance dispersal during colonization. *Heredity* 72: 312–317.

Ohta T. (1973) Slightly deleterious mutant substitutions in evolution. *Nature* 246: 96–98.

Pei C. (1934) The vascular plants of Nanking IV. *Contributions from the Biological Laboratory of the Science Society of China: Botanical Series* 9(2): 141–188.

Pons O. & Petit R. J. (1995) Estimation, variance and optimal sampling of gene diversity I. Haploid locus. *Theoretical and Applied Genetics* 90 (3–4): 462–472.

Pons O. & Petit R. J. (1996) Measuring and testing genetic differentiation with ordered versus unordered alleles. *Genetics* 144 (3): 1237–1245.

Pritchard J. K., Stephens M & Donnelly P. (2000) Inference of population structure using multilocus genotype data. *Genetics* 155(2): 945–959.

Quijano M., Iriondo J. M. & Torres E. (2012) Improving representativeness of genebank collections through species distribution models, gap analysis and ecogeographical maps. *Biodiversity and Conservation* 21 (1): 79–96.

R core team. (2019) *R: A language and environment for statistical computing*, R Foundation for Statistical Computing.

Ronquist F. & Huelsenbeck J. P. (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19: 1572–1574.

Rousset F. (1997) Genetic differentiation and estimation of gene flow from F-statistics under isolation by distance. *Genetics* 145 (4): 1219–1228.

Ruxton G. (2006) The unequal variance t-test is an underused alternative to Student's t-test and the Mann-Whitney U test. *Behavioral Ecology* 17 (4): 688–690.

Saff E. B. & Kuijlaars A. B. J. (1997) Distributing many points on a sphere. *Mathematical Intelligencer* 19 (1): 5–11.

Sagarin R. D. & Gaines S. D. (2002) The 'abundance centre' distribution: to what extent is it a biogeographical rule? *Ecological Letters* 5(1): 137–147.

Satterthwaite F. E. (1941) Synthesis of variance. *Psychometrika* 6 (5): 309–316.

Siebold P. F. von & Zuccarini J. G. (1846) Florae Japonpicae familiae naturales, adjectis generum et specierum exemplis selectis. Sectio altera. Plantae Dichotyledoneae (Gamopetalae, Monochlamydeae) et Monocotyledoneae. *Abhandlungen der Mathemat.-Physikalischen Classe der Königlich Bayerichen Akademie der Wissenschaften* 4 (3): 125–240.

Simmons M.P. & Ochoterena H. (2000) Gaps as characters in sequence-based phylogenetic analyses. *Systematic Biology* 49: 369–381.

Sokal R. R. & Rohlf F. J. (1995) *Biometry ed. 3*. W. H. Freeman and Co., New York.

Sokal R. R. & Rohlf F. J. (2012) *Biometry ed. 4*. W. H. Freeman and Co., New York.

Stamatakis A. (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30(9): 1312–1313.

Student. (1908) The probable error of a mean. *Biometrika* 6 (1): 1–25.

Suyama Y. & Matsuki Y. (2015) MIG-seq: an effective PCR-based method for genome-wide single-nucleotide polymorphism genotyping using the next-generation sequencing platform. *Scientific Reports* 5:16963.

Suzuki D. & Hikida T. (2011) Mitochondrial phylogeography of the Japanese pond turtle, *Mauremys japonica* (Testudines, Geoemydidae). *Journal of Zoological Systematics and Evolutionary Research* 49 (2): 141–147.

Takehana Y., Nagai N., Matsuda M., Tsuchiya K. & Sakaizumi M. (2003) Geographic variation and diversity of the cytochrome b gene in Japanese wild populations of

Medaka, *Oryzias latipes*. *Zoological Science* 20 (10): 1279–1291.

Tammes P. M. L. (1930) On the origin of number and arrangement of the places of exit on the surface of pollen grains. *Recueil des Travaux Botaniques Néerlandais* 27: 1–84.

Tamura K., Stecher G., Peterson D., Filipski A. & Kumar S. (2013) MEGA6: Molecular evolutionary genetics analysis version 6.0. *Molecular Biology and Evolution* 30 (12): 2725–2729.

Tanabe A.S. (2011) Kakusan4 and Aminosan: two programs for comparing nonpartitioned, proportional, and separate models for combined molecular phylogenetic analyses of multilocus sequence data. Molecular Ecology Resources 11: 914–921.

Tateishi Y. (2006) 5. Nanocnide Blume. *In*: Iwatsuki, K., Boufford, D. E. & Ohba, H. (Eds.) *Flora of Japan* IIa. Kodansha, Tokyo, pp. 89–90.

Thiers B. M. & Tulig M. (2019) Index Herbariorum. Website: http://sweetgum.nybg.org/science/ih/ (accessed on 9 November 2019).

Thompson S. K. & Seber G. A. F. (1966) Adaptive Sampling. Wiley, New York.

Tseng Y.-H., Monro A. K., Wei Y.-G. & Hu J.-M. (2019) Molecular phylogeny and morphology of *Elatostema s.l.* (Urticaceae): Implications for inter- and infrageneric classifications.

Turland N. J., Wiersema J. H., Barrie F. R., Greuter W., Hawksworth D. L., Herendeen P. S., Knapp S., Kusber W.-H., Li D.-Z., Marhold K., May T. W., McNeill J., Monro A. M., Prado J., Price M. J. & Smith G. F. (eds.) 2018: International Code of Nomenclature for algae, fungi, and plants (Shenzhen Code) adopted by the Nineteenth International Botanical Congress Shenzhen, China, July 2017. Regnum Vegetabile 159. Glashütten: Koeltz Botanical Books. (accessed: 12 December 2019).

Uwai S., Kogame K., Yoshida G., Kawai H. & Ajisaka T. (2009) Geographical genetic structure and phylogeography of the *Sargassum horneri*/*filicinum* complex in Japan, based on the mitochondrial cox3 haplotype. *Marine Biology* 156 (5): 901–911.

Viechtbauer W. (2007) Approximate confidence intervals for standardized effect sizes in the two-independent and two-dependent samples design. *Journal of Educational and Behavioral Statistics* 32(1): 39–60.

Viechtbauer W. (2010) Conducting meta-analysis in R with the metafor package. *Journal of Statistical Software* 36(3): 1–48.

Vogler A. J., Birdsell D., Price L. B., Bowers J. R., Beckstrom-Sternberg S. M., Auerbach R. K., Beckstorm-Sternberg J. S., Johansson A., Clare A., Buchhagen J. L., Petersen J. M., Pearson T., Vaissaire J., Dempsey M. P., Foxall P., Engelthaler D. M.,

Wagner D. M. & Keim P. (2009) Phylogeography of *Francisella tularensis*: Global expansion of a highly fit clone. *Journal of Bacteriology* 191 (8): 2474–2484.

Warton D. I., Duursma R. A., Falster D. S. & Taskinen S. (2012) smart 3- an R package for estimation and inference about allometric lines. *Methods in Ecology and Evolution* 3 (2): 257–259.

Wasserstein R., Lazar N. (2016) The ASA statement on p-values: context, process, and purpose. *The American Statistician* 70 (2): 129–133.

Weddell H. A. (1854) Revue de la famille des Urticées. *Annales des Sciences Naturelles; Botanique* série 4(1): 171–212.

Weddell H. A. (1869) Ordo CLXXXV. Urticaceae (1). *In*: Candolle, A. L. P. P. de (Ed.) *Prodromus systematis naturalis regni vegetabilis, sive enumeration contracta ordinum, generum, specierumque plantarum huc usque cognitarum, juxta methodi naturalis normas digesta*, 16 sectio prior. Lahure C., Paris, pp. 32–235.

Weiss G. H. & Kimura M. (1965) A mathematical analysis of the stepping stone model of genetic correlation. *Journal of Applied Probability* 2 (1): 129–149.

Welch B. L. (1938) The significance of the difference between two means when the population variances are unequal. *Biometrika* 29 (3): 350–362.

Welch B. L. (1947) The generalization of 'Student's' problem when several different population variances are involved. *Biometrika* 34 (1): 28–35.

White J. W., Rassweiler A., Samhouri J. F., Stier A. C. & White C. (2013) Ecologists should not use statistical significance tests to interpret simulation model results. *Oikos* 123 (4): 385–388.

White T. J., Bruns T., Lee S. & Taylor J. (1990) Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. *A Guide to Method and Applications* 18: 315–322.

Wu Z.-Y., Monro A. K., Milne R. I., Wang H., Yi, T.-S., Liu J. & Li D.-Z. (2013) Molecular phylogeny of the nettle family (Urticaceae) inferred from multiple loci of three genomes and extensive generic sampling. *Molecular phylogenetics and Evolution* 69: 814–827.

Yang Y.-P., Shih B.-L. & Liu H.-Y. (1996) 8. Urticaceae. *In*: Huang, T. (Ed.) *Flora of Taiwan second edition* 2. National Taiwan University, Taipei, pp.197–257.

Zar J. H. (2014) Biostatistical Analysis. Pearson Education Ltd., Edinburgh Gate.

Appendix

Supplementary figures

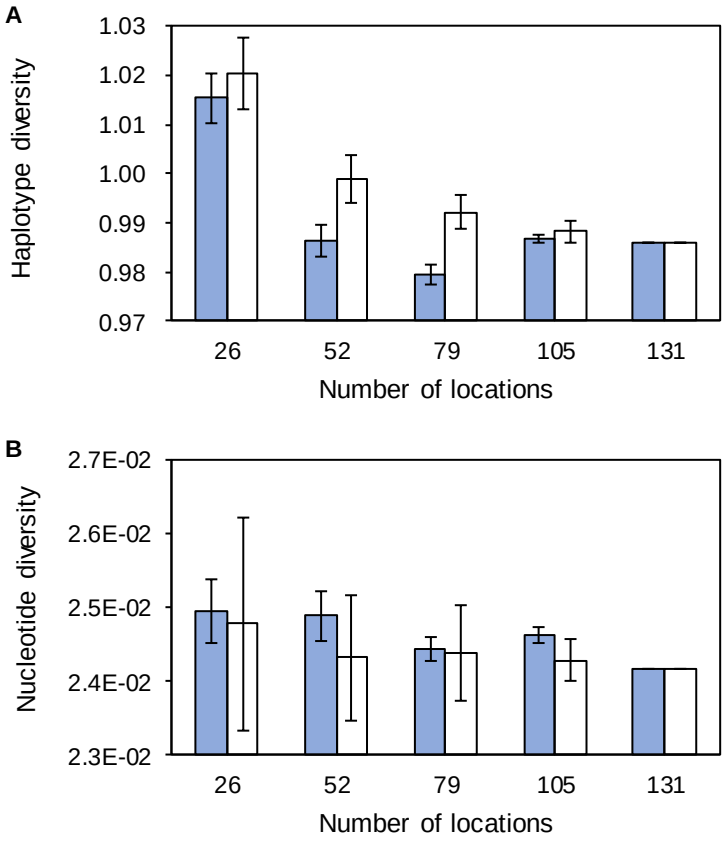

Figure S1. Diversity of *Macaca fuscata* samples using two sampling methods. A: haplotype diversity. B: nucleotide diversity. Blue: diversity of spatial sampling. White: that of random sampling. The original data are cited from (Kawamoto et al. 2007). Error bar: Standard deviation.

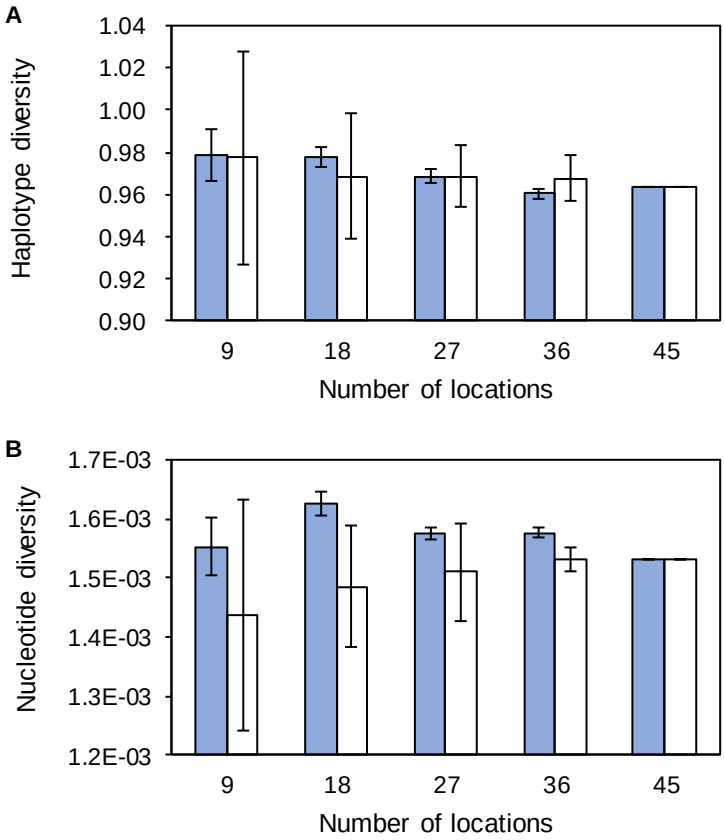

Figure S2. Diversity of *Fagus crenata* samples using two sampling methods. A: haplotype diversity. B: nucleotide diversity. Blue: diversity of spatial sampling. White: that of random sampling. The original data are cited from (Fujii et al. 2002). Error bar: Standard deviation.

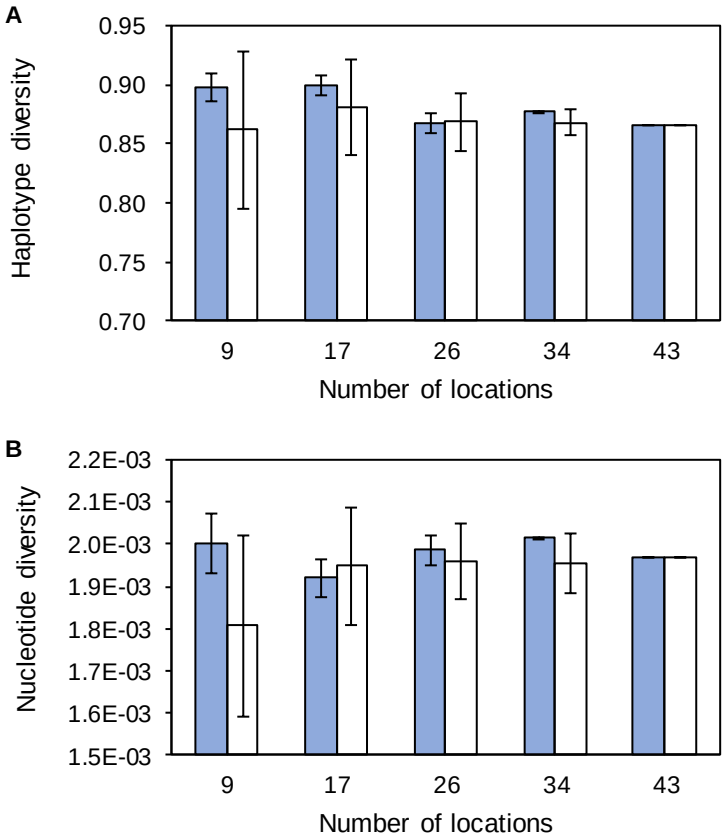

Figure S3. Diversity of *Mauremys japonica* samples using two sampling methods. A: haplotype diversity. B: nucleotide diversity. Blue: diversity of spatial sampling. White: that of random sampling. The original data are cited from (Suzuki & Hikida 2011). Error bar: Standard deviation.

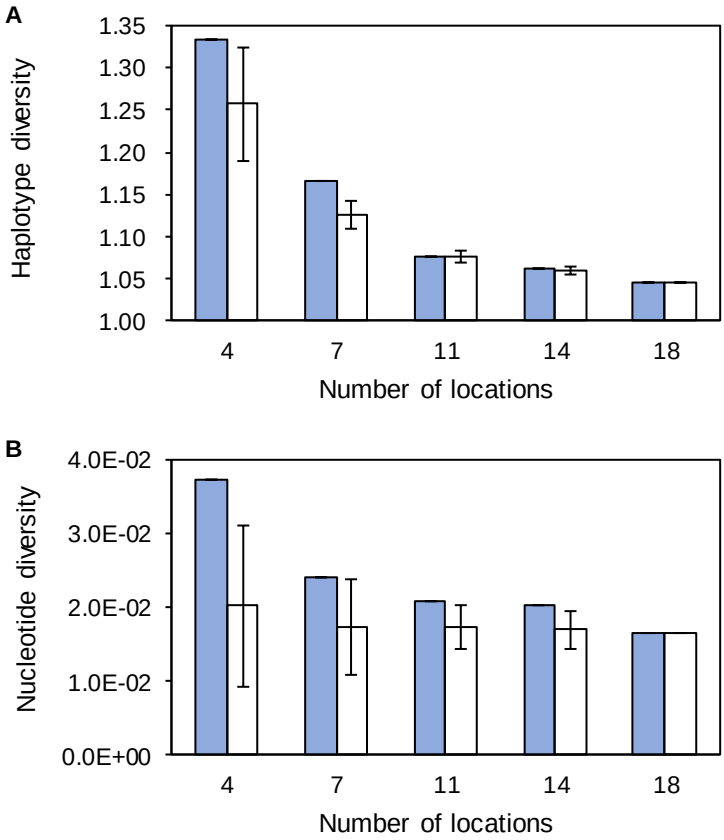

Figure S4. Diversity of *Oryzias sakizumii* (Clade A in (Takehana et al. 2003)) samples using two sampling methods. A: haplotype diversity. B: nucleotide diversity. Blue: diversity of spatial sampling. White: that of random sampling. The original data are cited from (Takehana et al. 2003). Error bar: Standard deviation.

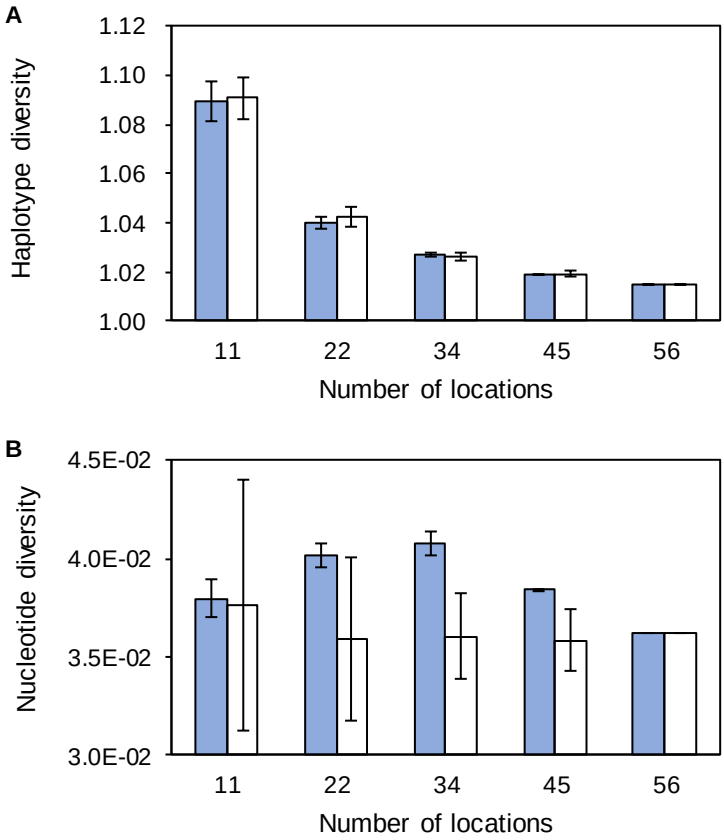

Figure S5. Diversity of *Oryzias latipes* (Clade B+C in (Takehana et al. 2003)) samples using two sampling methods. A: haplotype diversity. B: nucleotide diversity. Blue: diversity of spatial sampling. White: that of random sampling. The original data are cited from (Takehana et al. 2003). Error bar: Standard deviation.

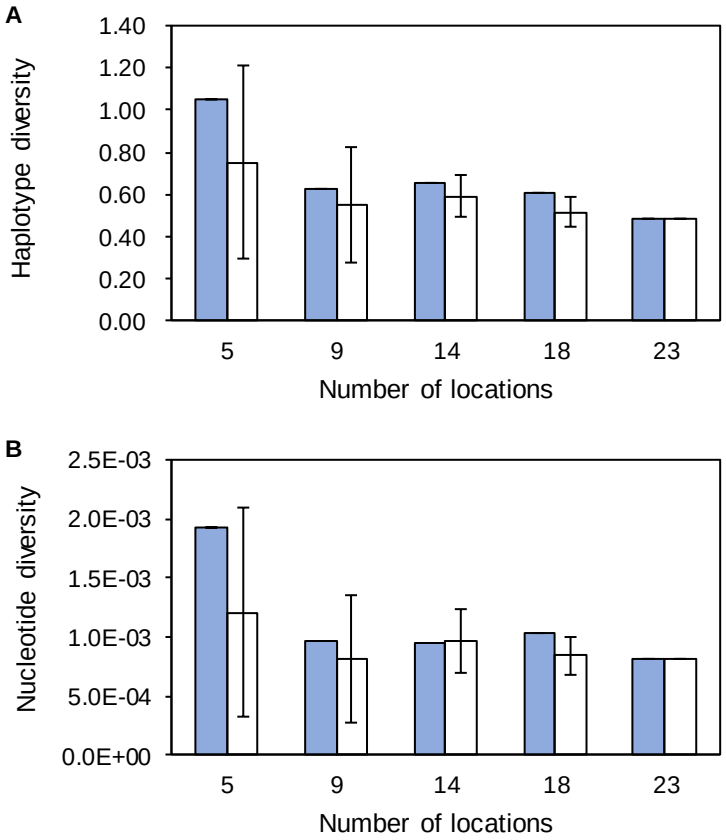

Figure S6. Diversity of *Parides alconius alconius* (including subsp. *yakushimanus*) samples using two sampling methods. A: haplotype diversity. B: nucleotide diversity. Blue: diversity of spatial sampling. White: that of random sampling. The original data are cited from (Kato & Yagi 2004). Error bar: Standard deviation.

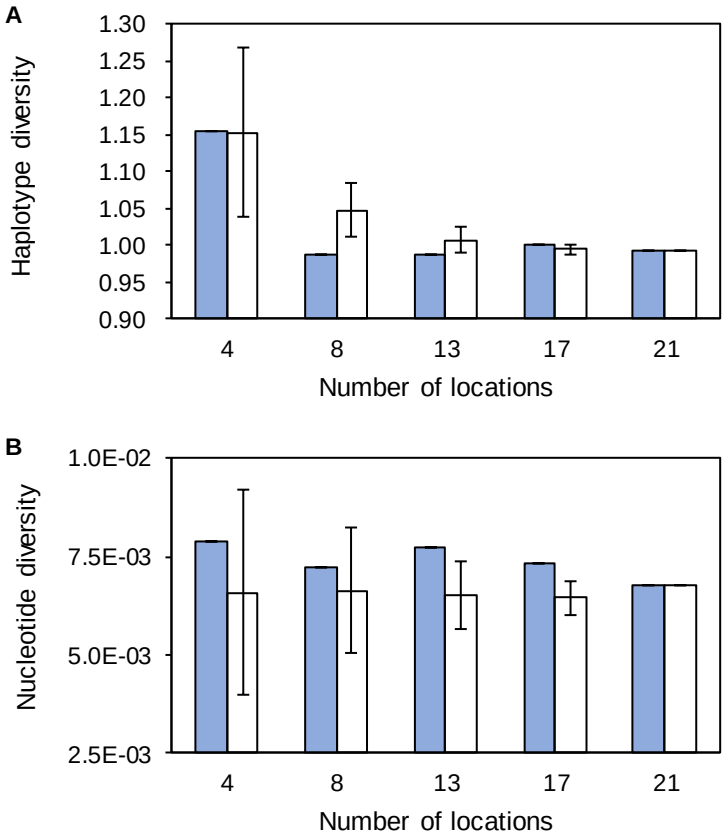

Figure S7. Diversity of *Amanita muscaria* Clade 1 in (Geml et al. 2008) samples using two sampling methods. A: haplotype diversity. B: nucleotide diversity. Blue: diversity of spatial sampling. White: that of random sampling. The original data are cited from (Geml et al. 2008). Error bar: Standard deviation.

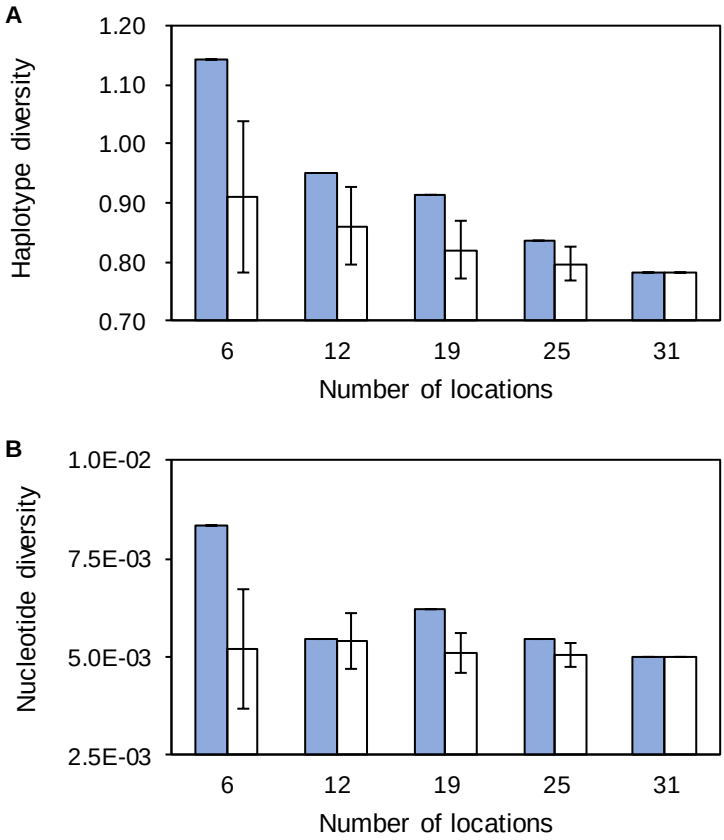

Figure S8. Diversity of *Amanita muscaria* Clade 2 in (Geml et al. 2008) samples using two sampling methods. A: haplotype diversity. B: nucleotide diversity. Blue: diversity of spatial sampling. White: that of random sampling. The original data are cited from (Geml et al. 2008). Error bar: Standard deviation.

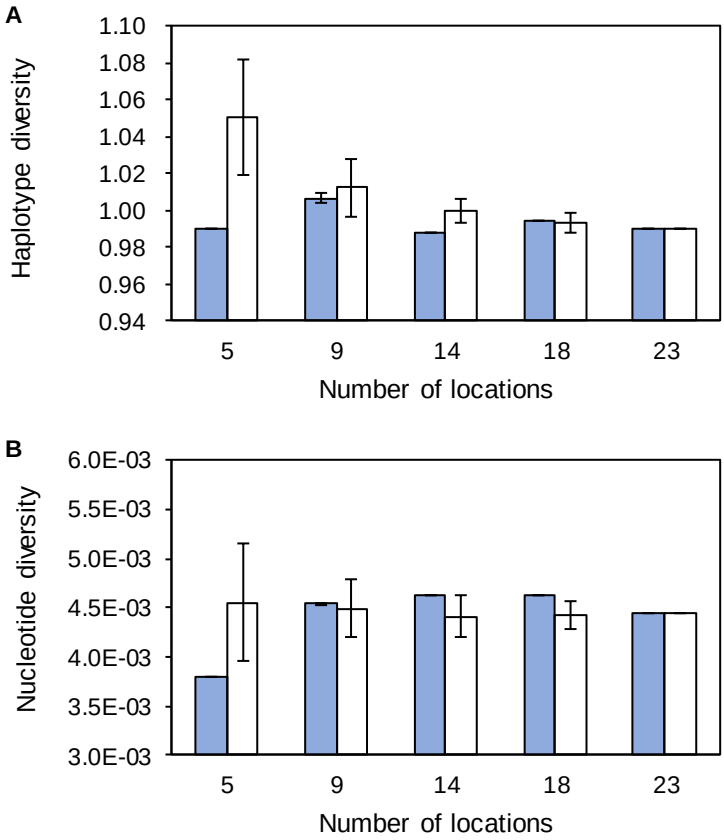

Figure S9. Diversity of *Ptionorhynchus violaceus violaceus* samples using two sampling methods. A: haplotype diversity. B: nucleotide diversity. Blue: diversity of spatial sampling. White: that of random sampling. The original data are cited from (Nicholls & Austin 2005). Error bar: Standard deviation.

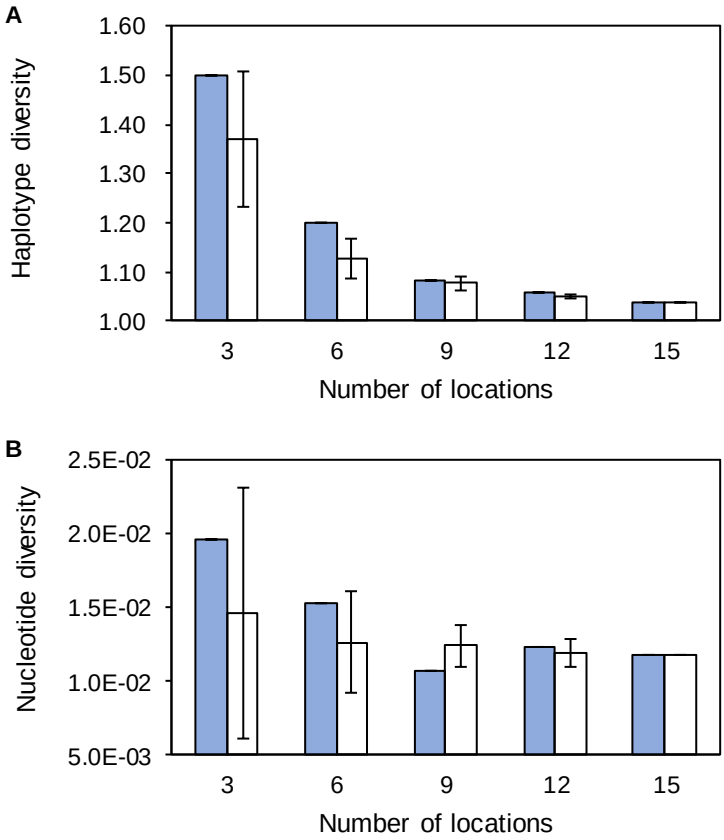

Figure S10. Diversity of *Wyeomyia smithii* samples using two sampling methods. A: haplotype diversity. B: nucleotide diversity. Blue: diversity of spatial sampling. White: that of random sampling. The original data are cited from (Emerson et al. 2010). Error bar: Standard deviation.

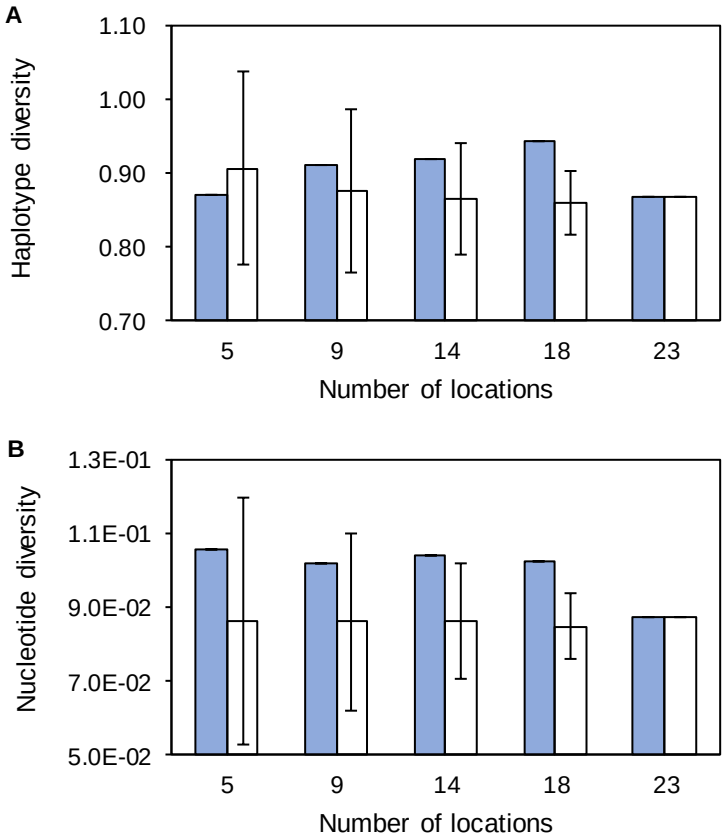

Figure S11. Diversity of *Francisella tularensis* subsp. *tularensis* samples using two sampling methods. A: haplotype diversity. B: nucleotide diversity. Blue: diversity of spatial sampling. White: that of random sampling. The original data are cited from (Vogler et al. 2009). Error bar: Standard deviation.

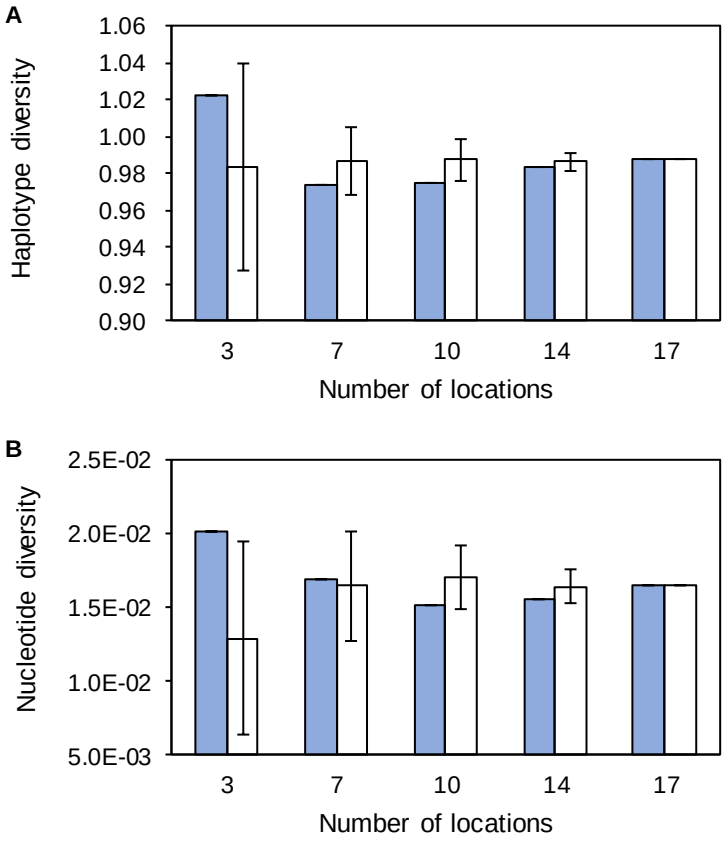

Figure S12. Diversity of *Echiniscoides sigismundi* SigiNorth in (Faurby et al. 2011) samples using two sampling methods. A: haplotype diversity. B: nucleotide diversity. Blue: diversity of spatial sampling. White: that of random sampling. The original data are cited from (Faurby et al. 2011). Error bar: Standard deviation.

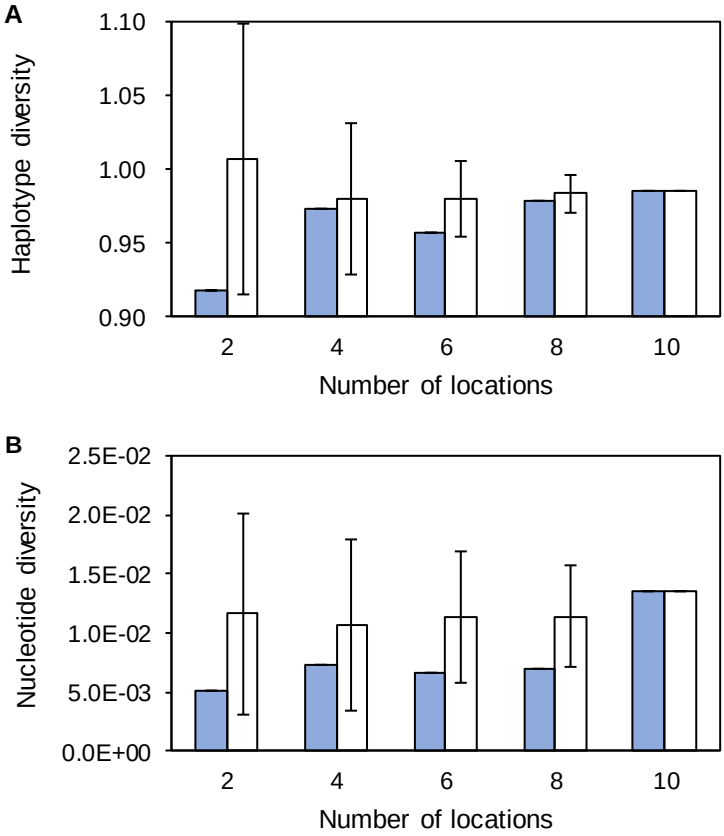

Figure S13. Diversity of *Echiniscoides sigismundi* SigiSouth in (Faurby et al. 2011) samples using two sampling methods. A: haplotype diversity. B: nucleotide diversity. Blue: diversity of spatial sampling. White: that of random sampling. The original data are cited from (Faurby et al. 2011). Error bar: Standard deviation.

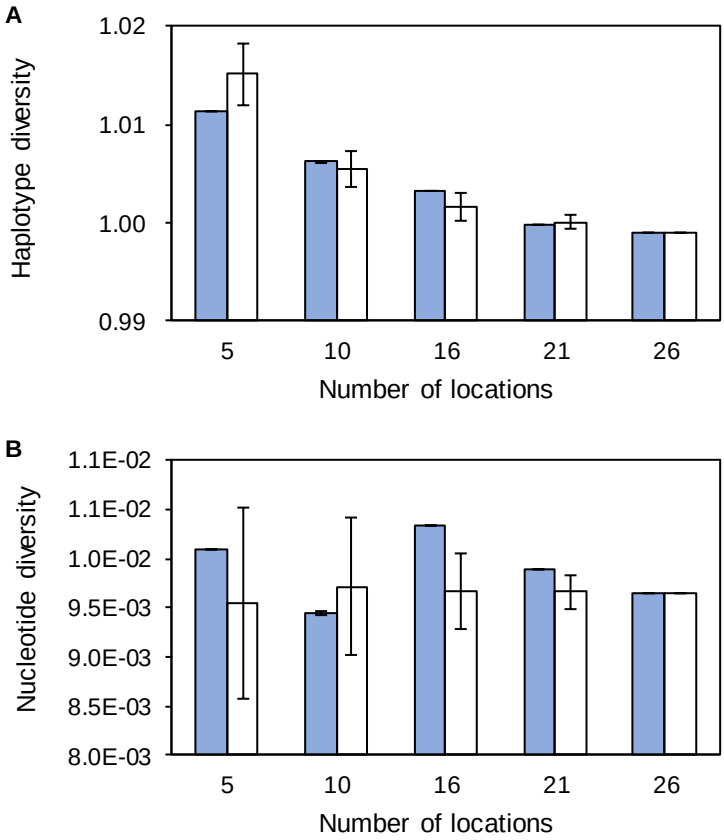

Figure S14. Diversity of *Paracentrotus lividus* samples using two sampling methods. A: haplotype diversity. B: nucleotide diversity. Blue: diversity of spatial sampling. White: that of random sampling. The original data are cited from (Maltagliati et al. 2010). Error bar: Standard deviation.

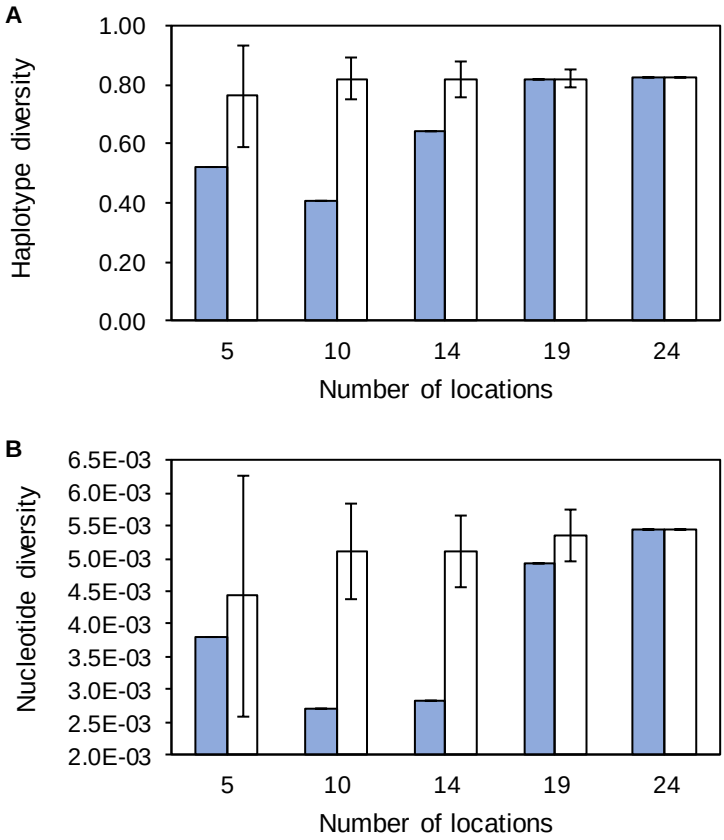

Figure S15. Diversity of *Sargassum horneri* samples using two sampling methods. A: haplotype diversity. B: nucleotide diversity. Blue: diversity of spatial sampling. White: that of random sampling. The original data are cited from (Uwai et al. 2009). Error bar: Standard deviation.

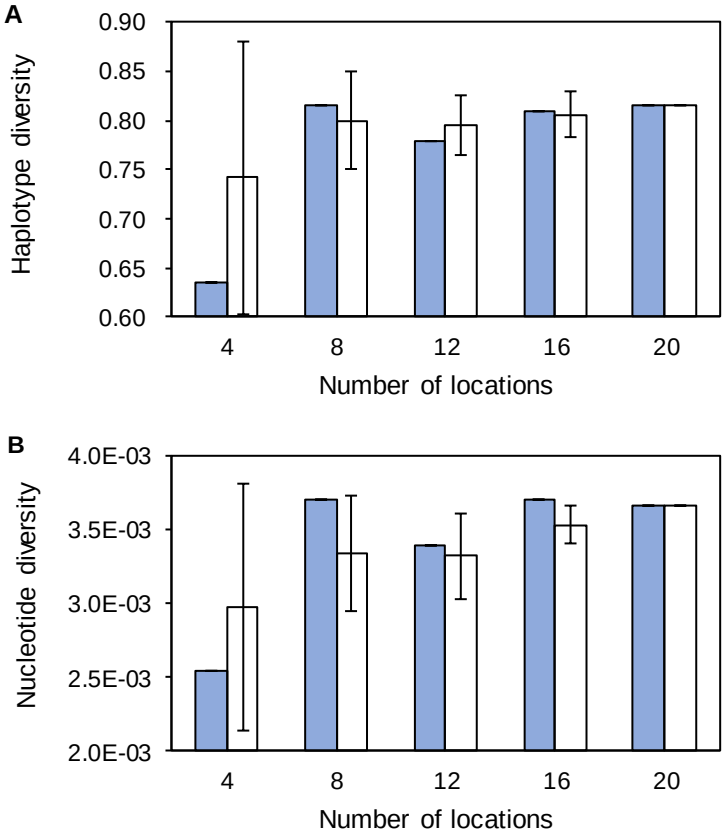

Figure S16. Diversity of *Gelidium lingulatum* samples using two sampling methods. A: haplotype diversity. B: nucleotide diversity. Blue: diversity of spatial sampling. White: that of random sampling. The original data are cited from (López et al. 2017). Error bar: Standard deviation.

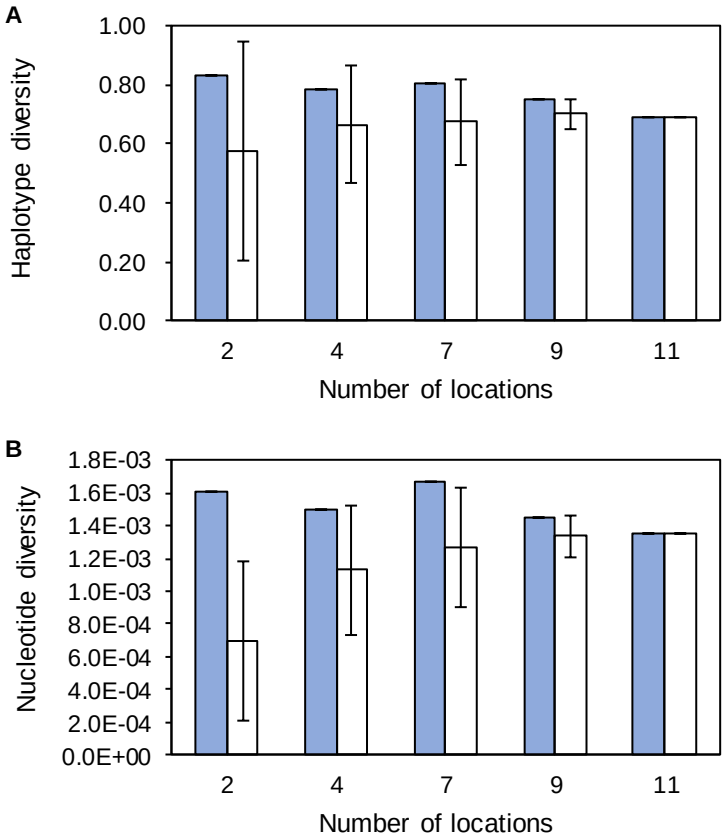

Figure S17. Diversity of *Gelidium rex* samples using two sampling methods. A: haplotype diversity. B: nucleotide diversity. Blue: diversity of spatial sampling. White: that of random sampling. The original data are cited from (López et al. 2017). Error bar: Standard deviation.

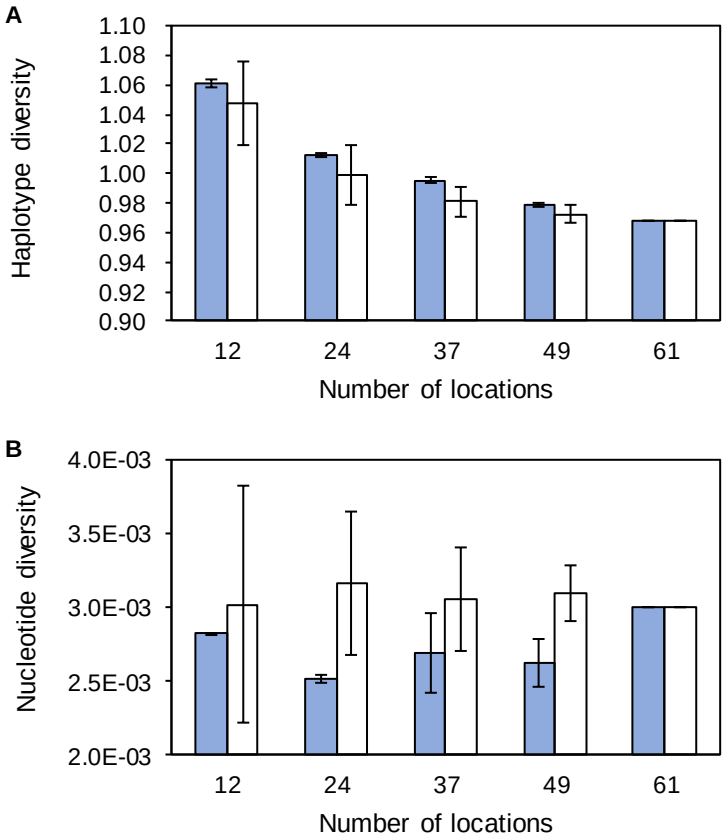

Figure S18. Diversity of *Pteridium aquilinum* samples using two sampling methods. The samples contain those of *P. caudatum*, which is a hybrid between *P. aquilinum* and another taxon. Its chloroplast sequence is included in the clade of *P. aquilinum*. A: haplotype diversity. B: nucleotide diversity. Blue: nucleotide diversity of spatial sampling. White: that of random sampling. The original data are cited from (Der et al. 2009). Error bar: Standard deviation.

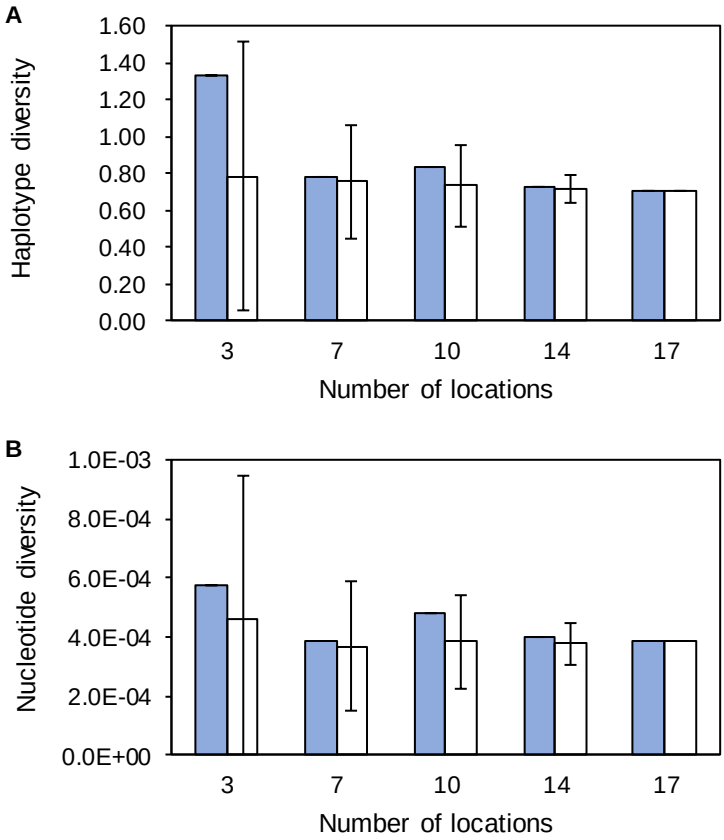

Figure S19. Diversity of *Pteridium aquilinum* subsp. *aquilinum* samples using two sampling methods. A: haplotype diversity. B: nucleotide diversity. Blue: diversity of spatial sampling. White: that of random sampling. The original data are cited from (Der et al. 2009). Error bar: Standard deviation.

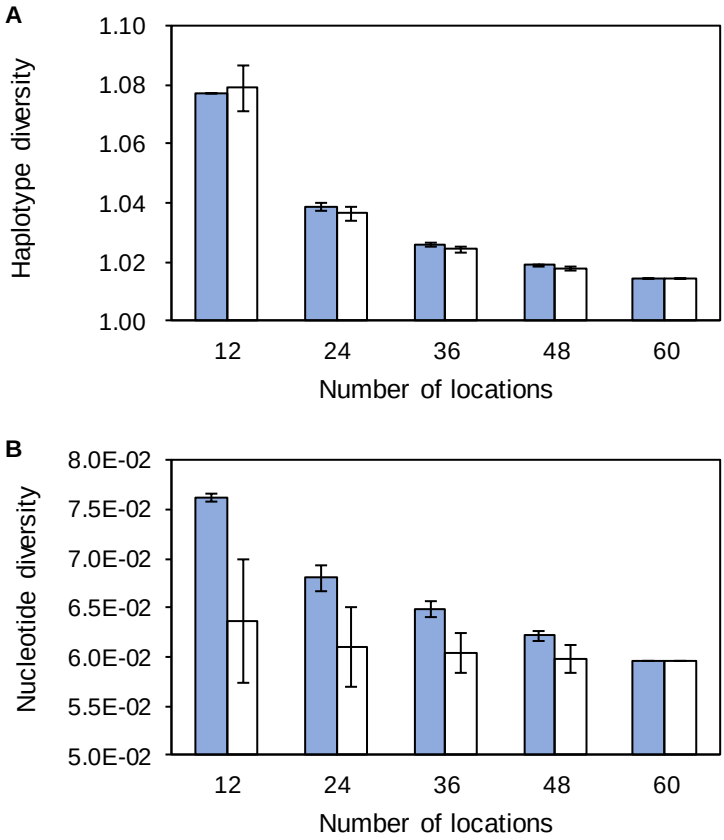

Figure S20. Diversity of *Metaphire sieboldi* samples using two sampling methods. A: haplotype diversity. B: nucleotide diversity. Blue: diversity of spatial sampling. White: that of random sampling. The original data are cited from (Minamiya et al. 2009). Error bar: Standard deviation.

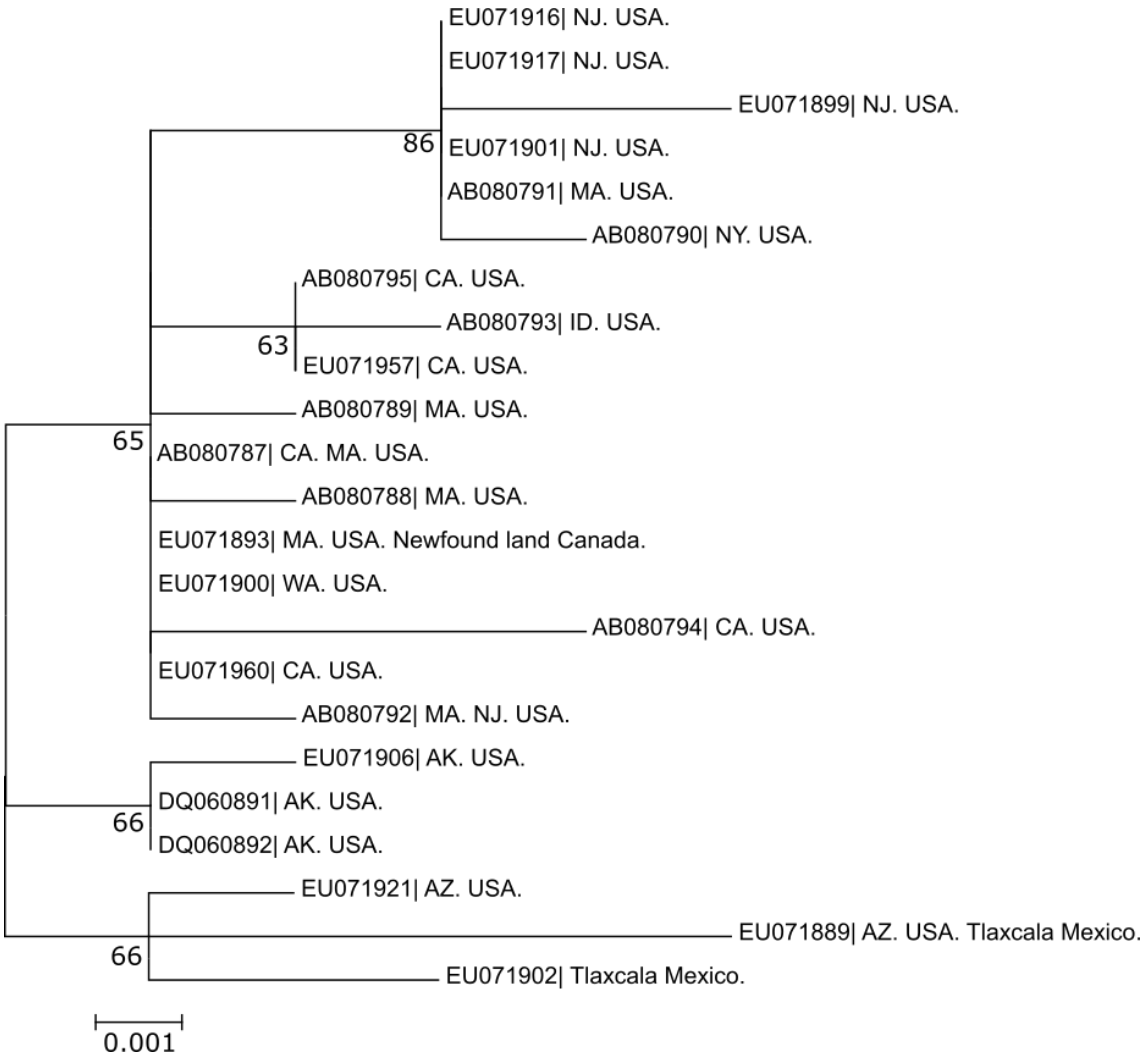

Figure S21. Phylogenetic tree of *Amanita muscaria* Clade 1 in (Geml et al. 2008). The tree was based on the internal transcribed spacer region using the maximum likelihood method in MEGA6 (Tamura et al. 2013). Completely identical sequences were collapsed into a single sequence. The confidence of branches was shown by nearby bootstrap values based on 1000 replicates. NJ: New Jersey. MA: Massachusetts. NY: New York. CA: California. WA: Washington. AK: Alaska. AZ: Arizona. The original data are available on TreeBASE.

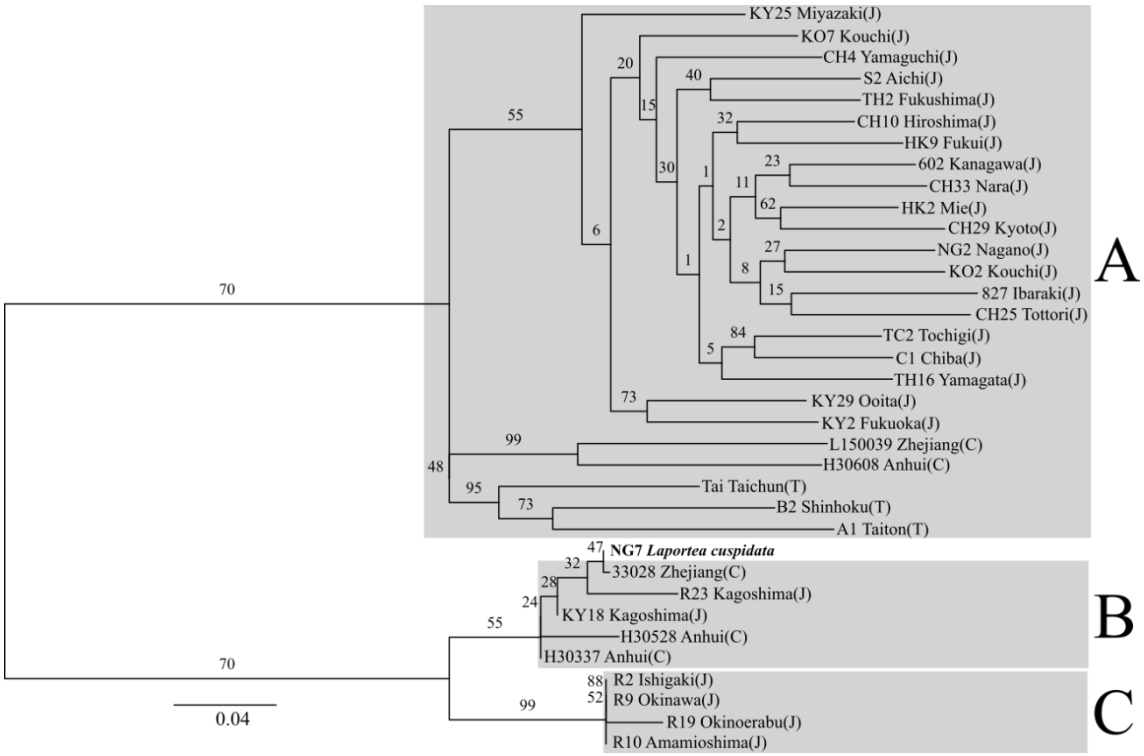

Figure S22. The ML tree based on MIG-seq including *L. cuspidata* (bold). Numbers near a branch stand for bootstrap values (100 times). Abbreviation after the sample locality shows the country of the locality. (C): Mainland China, (J): Japan, and (T): Taiwan.

Proof of unbiasedness and variation of e

In short, this proof is an application of the proof in Hedges (1981) to the statistic v in Welch (1938). Suppose two independently and normally distributed populations $N_1(\mu_1, \sigma_1^2)$ and $N_2(\mu_2, \sigma_2^2)$. Their sample means are $\overline{Y^1}$ and $\overline{Y^2}$, and their samples are Y_i^1 $(i = 1, \ldots, n_1)$ and Y_j^2 $(j = 1, \ldots, n_2)$. The statistic e^{biased} between them can be converted into

$$\sqrt{\tilde{n}}e^{biased} = \frac{\left(\overline{Y^1} - \overline{Y^2}\right)/\sqrt{(\sigma_1^2/n_1) + (\sigma_2^2/n_2)}}{\sqrt{wf/f}}, \tag{18}$$

where

$$\mathrm{w} = \frac{s_1^2/n_1 + s_2^2/n_2}{(\sigma_1^2/n_1) + (\sigma_2^2/n_2)}.$$

Here, since N_1 and N_2 are independently and normally distributed, the numerator of (18) has the normal distribution of $N(\theta, 1)$, where

$$\theta = \frac{\mu_1 - \mu_2}{\sqrt{(\sigma_1^2/n_1) + (\sigma_2^2/n_2)}},$$

and the s_i^2 is the same as (3). In the denominator, wf is approximately distributed as $\chi^2(f)$ (Welch 1938). Therefore, $\sqrt{\tilde{n}}e^{biased}$ is distributed as a non-central t variate with the non-centrality parameter θ and approximate degree of freedom f (16). From the nature of the non-central t distribution (Johnson & Welch 1940), the expected value of e^{biased} (12) is

$$\mathrm{E}\left(\sqrt{\tilde{n}}e^{biased}\right) = \theta\frac{\sqrt{f/2}\Gamma\{(f-1)/2\}}{\Gamma(f/2)}$$
$$\mathrm{E}\left(e^{biased}\right) = \theta/\sqrt{\tilde{n}}/J(f).$$

Now, supposing $\mathrm{r} = n_1/n_2$, then $\theta/\sqrt{\tilde{n}} = \epsilon_r$. In this case, the expected value of e (15) is

$$\begin{aligned}\mathrm{E(e)} &= \mathrm{E}\{e^{biased}J(f)\}\\ &= \mathrm{E}\left(e^{biased}\right)J(f)\\ &= \{\theta/\sqrt{\tilde{n}}/J(f)\}J(f)\\ &= \epsilon_r.\end{aligned}$$

Thus, e (15) is an unbiased estimator of ϵ_r (11). From the result of Johnson & Welch (1940), the variance of e^{biased} (12) is

$$\mathrm{var}\left(\sqrt{\tilde{n}}e^{biased}\right) = \frac{f}{f-2}(1+\theta^2) - \theta^2/J^2(f)$$

$$\mathrm{var}\left(e^{biased}\right) = \frac{f}{f-2}(1/\tilde{n} + \theta^2/\tilde{n}) - \theta^2/J^2(f)/\tilde{n}.$$

Therefore, the variation of e (15) is

$$\begin{aligned}\mathrm{var}(e) &= \mathrm{var}\{e^{biased}J(f)\}\\ &= \frac{f}{f-2}J^2(f)\{1/\tilde{n} + \left(\theta/\sqrt{\tilde{n}}\right)^2\} - \left(\theta/\sqrt{\tilde{n}}\right)^2\\ &= \frac{f}{f-2}J^2(f)(1/\tilde{n} + {\epsilon_r}^2) - {\epsilon_r}^2.\end{aligned}$$

■

Proof of unbiasedness and variation of c

The bias correction and derivation of the variance can be proved in the same way as that of d (5). The statistic c^{biased} (10) can be converted into

$$\sqrt{n-1}c^{biased} = \frac{\overline{Y^1} - C}{s/\sqrt{n_1 - 1}}, \tag{19}$$

and this (19) is distributed as a non-central t variate with non-centrality parameter

$$\frac{\mu - C}{\sigma/\sqrt{n_1 - 1}}$$

and degree of freedom $n_1 - 1$. Therefore, the expected value c^{biased} (10) is

$$\mathrm{E}\left(\sqrt{n_1 - 1}c^{biased}\right) = \frac{\mu - C}{\sigma/\sqrt{n_1 - 1}}\frac{\sqrt{(n_1-1)/1}\Gamma\{(n_1-1)/2\}}{\Gamma\{(n_1-1)/2\}}$$

$$\begin{aligned}\mathrm{E}\left(c^{biased}\right) &= \frac{\mu - C}{\sigma}\frac{1}{J(n_1-1)}\\ \mathrm{E}\left(c^{biased}\right) &= \gamma/J(n_1-1).\end{aligned}$$

Because $\mathrm{c} = c^{biased}J(n_1 - 1)$, the expected value of c (17) is

$$\begin{aligned}\mathrm{E}(c) &= \mathrm{E}\{c^{biased}J(n_1-1)\}\\ &= \mathrm{E}\left(c^{biased}\right)J(n_1-1)\\ &= \gamma.\end{aligned}$$

Thus, c is an unbiased estimator of the effect size parameter γ (9). The variance of c^{biased} (10) is

$$\mathrm{var}\left(\sqrt{n_1 - 1}c^{biased}\right) = \frac{n_1 - 1}{n_1 - 3}\{1 + \left(\frac{\mu - C}{\sigma/\sqrt{n_1 - 1}}\right)^2\} - \left(\frac{\mu - C}{\sigma/\sqrt{n_1 - 1}}\right)^2\frac{1}{J(n_1-1)}$$

$$\begin{aligned}\mathrm{var}\left(c^{biased}\right) &= \frac{n_1 - 1}{n_1 - 3}\{\frac{1}{n_1 - 1} + \left(\frac{\mu - C}{\sigma}\right)^2\} - \left(\frac{\mu - C}{\sigma}\right)^2\frac{1}{J(n_1-1)}\\ \mathrm{var}\left(c^{biased}\right) &= \frac{n_1 - 1}{n_1 - 3}\left(\frac{1}{n_1 - 1} + \gamma^2\right) - \gamma^2\frac{1}{\mathrm{J}^2(n_1-1)}.\end{aligned}$$

Therefore, the variation of c (17) is

$$\begin{aligned}\mathrm{var}(c) &= \mathrm{var}\{c^{biased}J(n_1-1)\}\\ &= \mathrm{var}(c^{biased})J^2(n_1-1)\\ &= \frac{n_1-1}{n_1-3}J^2(n_1-1)\left(\frac{1}{n_1-1}+\gamma^2\right)-\gamma^2.\end{aligned}$$

■

Proof of consistency of c

First, I treat the proof of c which is simpler than that of e. For the proof, a lemma must be introduced.

Lemma 1. Assume random samples $Y_1^1, \dots, Y_n^1$ from the population with the population mean μ_1 and the population variance σ_1^2, and consider a parameter β and its statistic $b = b(Y_1^1, \dots, Y_n^1)$. Then,

$$\left[\mathrm{b\ is\ an\ unbiased\ estimator\ of}\ \beta, \mathrm{and}\ \lim_{n\to\infty}\mathrm{var(b)} \to 0\right]$$

$$\Rightarrow [\mathrm{b\ is\ a\ consistent\ estimator\ of}\ \beta.]$$

Proof.

$$\begin{aligned}\mathrm{E}(|\mathrm{b}-\beta|^2) &= \mathrm{E}(b-\beta)^2 + \mathrm{var}(\mathrm{b}-\beta)\\ &= \{\mathrm{E(b)}-\mathrm{E}(\beta)\}^2 + \mathrm{var(b)}\\ &= \{\mathrm{E(b)}-\beta\}^2 + \mathrm{var(b)}\end{aligned}$$

Given $\mathrm{E(b)} = \beta$ and $\lim_{n\to\infty}\mathrm{var(b)} \to 0$,

$$\lim_{n\to\infty}[\{\mathrm{E(b)}-\beta\}^2 + \mathrm{var(b)}] \to 0.$$

Therefore, b is a mean square consistent estimator of β, namely,

$$\lim_{n\to\infty}\mathrm{E}(|\mathrm{b}-\beta|^2) \to 0.$$

Here, for an arbitrary positive number ε, applying Chebyshev's inequality (Chebyshev 1867) provides

$$\begin{aligned}\mathrm{P}(|\mathrm{b}-\beta| \geq \varepsilon) &= \mathrm{P}(|\mathrm{b}-\beta|^2 \geq \varepsilon^2)\\ &\leq \mathrm{E}(|\mathrm{b}-\beta|^2)/\,\varepsilon^2.\end{aligned}$$

The result sown above provides

$$\lim_{n\to\infty}\mathrm{E}(|\mathrm{b}-\beta|^2)/\varepsilon^2 \to 0.$$

Therefore, using the squeeze theorem, it shows

$$\lim_{n\to\infty}\mathrm{P}(|\mathrm{b}-\beta| \geq \varepsilon) \to 0.$$

Thus, b is a consistent estimator of β.

■

Now, let's move on to the proof about c (17). When $n_1 \to \infty$, the variance of c will be

$$\begin{aligned}\lim_{n_1\to\infty} \mathrm{var}(c) &= \lim_{n_1\to\infty} \left\{\frac{n_1-1}{n_1-3} J^2(n_1-1)\left(\frac{1}{n_1-1}+\gamma^2\right)-\gamma^2\right\} \\ &\to 1\cdot J^2(\infty)(1/\infty+\gamma^2)-\gamma^2 \\ &= 0.\end{aligned}$$

Hence, $\lim_{n_1\to\infty} \mathrm{var}(c) \to 0$, and c is an unbiased estimator of γ. Therefore, based on lemma 1, c (17) is a consistent estimator of γ (9).

■

Proof of consistency of e

On the other hand, e (15) is consisted of two populations. Therefore, a variation of the previous lemma is necessary.

Lemma 2. Assume two random samples $Y_1^1, \dots, Y_{n_1}^1$ and $Y_1^2, \dots, Y_{n_2}^2$ from the populations with the population means μ_1 and μ_2, and the population variances σ_1^2 and σ_2^2, respectively. Consider a parameter β and its statistic $b = b\left(Y_1^1, \dots, Y_{n_1}^1; Y_1^2, \dots, Y_{n_2}^2\right)$. Then,

$$\left[\text{b is an unbiased estimator of } \beta, \text{and } \lim_{(n_1,n_2)\to(\infty,\infty)} \mathrm{var(b)} \to 0\right]$$
$$\Rightarrow [\text{b is a consistent estimator of } \beta.]$$

This lemma can be proved in the same way as lemma 1.

Now, consider $n_1 = \mathrm{r}\Phi$ and $n_2 = \Phi$, to think $\Phi \to \infty$, which equals to $(n_1, n_2) \to (\infty, \infty)$. Note that $\mathrm{r} > 0$ and $\theta > 0$, since $n_1 \geq 1$ and $n_2 \geq 1$. Using r and Φ, f (6) and $\tilde{n}$ (14) can be expressed as

$$f = \frac{(s_1^2/r + s_2^2)^2}{s_1^4/\{r^2(\mathrm{r}\Phi-1)\} + s_2^4/\{(1/r)^2(\Phi-1)\}}$$

and

$$\tilde{n} = \mathrm{r}\Phi/(r+1).$$

Therefore, when $\Phi \to \infty$, the variance of e (15) will be

$$\lim_{\Phi\to\infty} \mathrm{var}(e) = \lim_{\Phi\to\infty} \left\{\frac{f}{f-2} J^2(f)(1/\tilde{n} + {\epsilon_r}^2) - {\epsilon_r}^2\right\}$$

$$= \lim_{\Phi\to\infty} \left\{\frac{1}{1-2/f} J^2(f)(1/\tilde{n} + {\epsilon_r}^2) - {\epsilon_r}^2\right\}$$

$$\rightarrow \frac{1}{1-2/\infty} J^2(\infty)(1/\infty + {\epsilon_r}^2) - {\epsilon_r}^2$$

$$= \frac{1}{1-0} \cdot 1 \cdot (0 + {\epsilon_r}^2) - {\epsilon_r}^2$$

$$= 0.$$

The limit does not contain r, meaning $\lim_{(n_1,n_2)\to(\infty,\infty)} \mathrm{var(e)}$ always gives an identical value 0. Also, e is an unbiased estimator ϵ_r (11). Therefore, based on lemma 2, e (15) is a consistent estimator of ϵ_r (11).

■

Simulation source code

The simulation study in Chapter 3 was conducted using the following code in R:

```
library(es.dif)
library(Matrix)
library(metafor)

rep<-100000
n_sd<-10
sampleSize<-matrix(0,nrow=9,ncol=2)
sampleSize[1,]<-c(10,10)
sampleSize[2,]<-c(10,20)
sampleSize[3,]<-c(10,30)
sampleSize[4,]<-c(20,10)
sampleSize[5,]<-c(20,20)
sampleSize[6,]<-c(20,30)
sampleSize[7,]<-c(30,10)
sampleSize[8,]<-c(30,20)
sampleSize[9,]<-c(30,30)

d<-numeric(rep)
var_d<-numeric(rep)
ci_lb_d<-numeric(rep)
ci_ub_d<-numeric(rep)

e<-numeric(rep)
```

```
var_e<-numeric(rep)
ci_lb_e<-numeric(rep)
ci_ub_e<-numeric(rep)

Bonett<-numeric(rep)
var_Bonett<-numeric(rep)
ci_lb_Bonett<-numeric(rep)
ci_ub_Bonett<-numeric(rep)

result_d<-matrix(0,nrow=(n_sd)*nrow(sampleSize),ncol=4)
result_e<-matrix(0,nrow=(n_sd)*nrow(sampleSize),ncol=4)
result_Bonett<-matrix(0,nrow=(n_sd)*nrow(sampleSize),ncol=4)
counter <-1
for(k in 1:nrow(sampleSize))
{
        for(j in 1:n_sd)
        {
                for(i in 1:rep)
                {
                        data1<-rnorm(sampleSize[k,1],1,1)
                        data2<-rnorm(sampleSize[k,2],0,j)

                        temp_d <-es.d(data1,data2,vector_out=T)
                        d[i] <- temp_d[1]
                        var_d[i]<- temp_d[2]
                        ci_lb_d[i]<-temp_d[3]
                        ci_ub_d[i]<-temp_d[4]

                        temp_e <-es.e(data1,data2,vector_out=T)
                        e[i] <- temp_e[1]
                        var_e[i]<- temp_e[2]
                        ci_lb_e[i]<-temp_e[3]
                        ci_ub_e[i]<-temp_e[4]

                        temp_Bonett  <-summary(escalc(measure="SMDH",m1i=me
an(data1),m2i=mean(data2),sd1i=sd(data1),sd2i=sd(data2),n1i=length(data1),n2i=lengt
```

```
h(data2)))
                        Bonett[i] <- temp_Bonett[1,1]
                        var_Bonett[i]<- temp_Bonett[1,2]
                        ci_lb_Bonett[i]<-temp_Bonett[1,5]
                        ci_ub_Bonett[i]<-temp_Bonett[1,6]

                }
                result_d[counter,]<-c(mean(d),mean(var_d),mean(ci_lb_d),mean(ci_u
b_d))
                result_e[counter,]<-c(mean(e),mean(var_e),mean(ci_lb_e),mean(ci_u
b_e))
                result_Bonett[counter,]<-  c(mean(Bonett),mean(var_Bonett),mean(ci
_lb_Bonett),mean(ci_ub_Bonett))
                counter <- counter + 1
        }
}
```