Supporting information for

Waves out of the Korean Peninsula and inter- and intra-species replacements in freshwater fishes in Japan

Shoji Taniguchi,[1] Johanna Bertl,[2] Andreas Futschik,[3]
Hirohisa Kishino,[1] and Toshio Okazaki[1*]

[1]Graduate School of Agricultural and Life Sciences, The University of Tokyo, Japan
[2]Department of Mathematics, Aarhus University, Denmark
[3]Department of Applied Statistics, Johannes Kepler University Linz, Austria
[*]Corresponding author

S1 Supplementary document

1. PCR and sequencing of ND II and cytochrome b

1.1 PCR

Copy DNA was amplified by a Perkin Elmer Cetus (Irvine , CA) DNA thermal cycler under the following conditions: DNA denatured at 92°C for 40 s; primers annealed at 48~52°C (changed by species) for 60 s; copy DNA extension at 72°C for 120 s, for 28~30 cycles. Restriction endonucleases were purchased from New England Biolabs (Berverly, MA), Amersham International plc (Amersham, U.K.), or Takara (Shiga, Jpn) and used according to the manufacturer's instructions.

Hemibarbus longirostris: Following Hall and Nawrocki (1995), we carried out PCR on the ND1-16SRNA region of mtDNA (about 2.0 Kbp) using the following primers.
Forward: 5′-ACCCCGCCTGTTTACCAAAAACAT-3′
Reverse: 5′-GGTATGAGCCCGATAGCTTA-3′

Fifteen types of restriction enzymes were utilized: AciI, AfaI, AluI, BfaI, BstUI, DdeI, HaeIII, HhaI, HinfI, MboI, MspI, NlaIII, ScrFI, Sau96I, and TaqI.

Nipponocypris temminckii: Following Hall and Nawrocki (1995), the same primer as that for *H. longirostris* was used. Thirteen types of restriction enzymes were utilized: AfaI, AluI, BstUI, DdeI, HaeIII, HhaI, HinfI, MboI, MspI, NlaIII, ScrFI, Sau96I, and TaqI. From the cleavage type of each enzyme and the results of sequence analysis, it was found that individuals in clades F and G can be distinguished by differences in the cleavage type of BstUI, DdeI and TaqI.

Carassius spp.: Following Hall and Nawrocki (1995), the same primer as that for *H. longirostris* was used. Ten types of restriction enzymes were used: AfaI, BfaI, BstUI, HhaI, HinfI, MboI, MspI, NlaIII, ScrFI, and TaqI were adopted. The cleavage patterns of six enzymes could discriminate *C. cuvieri* from other *Carassius* species.

Tanakia limbata, *T. koreensis* and its related species: following Palumbi et al. (1991), we performed PCR on about 2.2 Kbp, including the control-12 SRNA regions of mtDNA, using the following primers.
Forward: Cb3R-L: 5′-CATATTAAACCCGAATGATATTT-3′
Reverse: 12SAR-H: 5′-ATAGTGGGGTATCTAATCCCAGTT-3′

Fifteen types of restriction enzymes were utilized: AciI, AfaI, AluI, BfaI, BstUI, DdeI, HaeIII, HhaI, HinfI, MboI, MspI, NlaIII, ScrFI, Sau96I, and TaqI.

1.2 Sequencing

First, total DNA was extracted from white muscle tissue (Asahida, Kobayashi, Taitoh, and Nakayama, 1996). Next, a partial region of the mitochondrial gene ND II was amplified by PCR using the following primer pairs designed based on the mtDNA sequence of *Cyprinus carpio* (Chang, Huang, and Lo, 1994): (5′-TWTYGGGCCCATACCCCRAA-3′) and (5′-GCTTTGAAGGCTYTTRGTCT-3′). PCR was conducted for 30 cycles at 94°C for 1 min, 52°C for 1 min, and 72°C for 2 min. The amplified DNA product was purified with a QIA quick PCR Purification Kit (Qiagen, Germany), and sequences were determined by an automated DNA sequencer (Applied Biosystem 377A). Cytochrome b sequence data were obtained in the same way. Primers included L14724 (Palumbi et al., 1991) (5′-TGACTTGAARAACCAYCGYYG-3′) and H15915 (Aoyama, Watanabe, Ishikawa, Nishida, and Tsukamoto, 2000) (5′- ACCTCCGATCTYCGGATTACAAGAC-3′). Copy DNA was amplified by a Perkin Elmer Cetus (Irvine, CA) DNA thermal cycler under the following conditions: DNA denatured at 92°C for 40 s; primers annealed at 48~50°C for 60 s; copy DNA extension at 72°C for 120 s, for 30 cycles. Multiple alignment of the nucleotide sequences was performed with software CLUSTRAL (Higgins and Sharp, 1988) and subsequently adjusted by eye.

2. Divergence time estimation

To estimate the evolutionary rate, 22 sequences were selected from each clade. The selected sequences were as follows.

Table S1

Correspondence of the sequence name between ND2 and cytochrome b.

ND2	Cytochrome b	clade
21_KAWAMUTSU	1_KOUZUKI	G
30_KAWAMUTSU	5_SHIMADA	F
47_KAWAMUTSU	10_TOKUEK	G
52_KAWAMUTSU	18_KYUUKA	B
57_KAWAMUTSU	13_SANTAB	C
60_KAWAMUTSU	9_KYUUKAW	G
64_KAWAMUTSU	11_DOUHUK	E
66_KAWAMUTSU	16_CHOSEN	A
68_KAWAMUTSU	15_RYUUDE	A
77_KUMAGAWA	12_KUMAGA	D
85_KONYOU	14_KONYOU	C
87_NISHIGAMI2	17_NISHIK	B
105_TAKAYAMA2	23_TAKAYA	C
106_KUGUNO	19_KUGUNO	C
115_NISHIKI	6_NISHIKI	G

124_OOTA	2_TOGOUCH	G
137_NAKA	4_NAKA	G
140_MIYA	8_MIYA	C
182_GUNKE	3_GUNNKE	F
210_MIYAKODA	7_MIYAKOD	C
211_SUMIYOSHI	21_SUMIYO	F
212_MACHINO	22_MACHIN	F

The topology of the Maximum Likelihood (ML) tree and sequence data were imported into MCMCTREE package. In MCMCTREE analysis, we assumed HKY85 as the model of nucleotide substitution and the *correlated rate* as the model of evolutionary rate. Since no reliable fossil record was available to set calibration points, we estimated relative age by setting the time at the tree root as 1. Additionally, we defined a few loose calibrations based on ML and Bayesian trees (Figure B2a, B2b) to prevent poor mixing of MCMC due to undue deviation from the reconstructed phylogenetic trees, which would not influence the order of node 1, 2 or 3. That is, we set four calibrations (Figure B3), node 4: ' < 0.5 ', node 5: ' < 0.8 ', node 6: ' < 0.75 ', and node 7: ' > 0.99 < '. Default values were adopted for other hyper-parameters. We performed MCMC simulation, and sampled the parameter set every 2000 iterations. By excluding the first 2000 parameter sets as burn-in, we obtained the final MCMC sample of 20,000 parameter sets. We estimated the posterior distribution of the difference between the relative ancestral age of clade C and F, as well as clade C and G, by calculating differences in the MCMC sample. Given the migration of clade F as 1.31 Ma, the conditional posterior of the migration period of clade G was estimated from MCMC samples obtained with BEAST and MCMCTREE. We calculated the ratio between migration times of clades G and F for each sample and then multiplied them by 1.31.

3. Detailed information for the estimation of the ancestral distribution, BayArea

Table S2

Location of the 21 areas throughout Korea and Japan.

Name	longitude	latitude
Kyushu-Southeast	131.4	31.9
Kyushu-Southwest	130.6	32.5
Kyushu-Northwest	130.7	33.7
Kyushu-Northeast	131.6	33.2

Shikoku-South	133.4	33.5
Shikoku-North	134.1	34.2
Chugoku-Southwest	132.5	34.5
Chugoku-Southeast	133.9	34.7
Chugoku-North	132.3	35
Kinki-Middle	135.5	34.7
Kinki-North	135.1	35.3
Kinki-South	135.2	34.2
Tokai-Ise Bay	136.7	35.2
Tokaik-East	137.8	34.8
Hokuriku-West	136.8	36.6
Han-Riv	127.3	37.5
Geum-Riv	127.4	36.5
Yeongsan-Riv	126.6	34.8
Seomjin-Riv	127.1	35
Nakdong-Riv	128.3	36.2
Yeongdong	129.3	37.2

3. Simulated distribution-formation process of *Nipponocypris temminckii*

We simulated the formation process of biogeographic distribution by generating the dynamics of the states on the 329 evenly distributed lattice-like grids. Their envelope covered the whole range in distribution of *N. temminckii* throughout Japan, expanded to the estimated coastal line at the time of glaciation periods (120 m below the current sea level; Figure 2) (Fairbanks, 1989; Rohling et al., 1998). The distance between points was defined as the geographic distance. Distances were calculated using the R (R Core Team, 2017) package geosphere (Karney, 2013).

As reported in our Results, phylogeographic analysis implied a specific formation scenario. That is, clade C arrived first, and clades F and G, in turn, migrated into Japan from the Korean Peninsula. We assumed that clade F had migrated into Western Japan 1.31 Ma (Figure B2b). The state of the simulation was the clade assignment of each grid. The simulation started with the state of assignment to clade C, except for the three points at northern Kyushu, which were assigned to clade F. In *r* Ma, clade G migrated. At each simulation step, the clade at each point had multiple

offspring. One stayed at the same point and the others dispersed to nearby points. We assumed the distance (x) that an offspring dispersed within time t followed a gamma distribution, in accordance with:

$$Gamma(shape = t \cdot m/s\,, scale = s) \tag{1}$$

where m is the expected distance that an offspring disperses per unit time (km/Kyr), and s is the scale parameter. In this model, the mean dispersal distance in time t is tm, and the variance is $tm{\cdot}s$. The probability that an offspring disperses from point i to point j is

$$p_{ij} = P\left(X > d_{ij}\right) \tag{2}$$

where d_{ij} is the distance between points i and j. If more than one clade coexists at point i as a result of dispersal, one clade was chosen randomly with a probability that reflects the difference in the fitness between clades. As a simple model representing the fitness difference between clades, we assumed that both the selective advantage of clade F over clade C, and that of clade G over clade F, was α, and that the selective advantage of clade G over clade C, was α^2. With this model, the replacement probability in a case of clade-coexistence is:

$$\begin{aligned}
&p(F|C,F) = p(G|F,G) = \alpha,\\
&p(C|C,F) = p(F|F,G) = 1-\alpha,\\
&p(G|C,G) = \alpha^2/\{\alpha^2 + (1-\alpha)^2\},\\
&p(C|C,G) = (1-\alpha)^2/\{\alpha^2 + (1-\alpha)^2\},\\
&p(C|C,F,G) = (1-\alpha)^2/\{\alpha^2 + \alpha(1-\alpha) + (1-\alpha)^2\},\\
&p(F|C,F,G) = \alpha(1-\alpha)/\{\alpha^2 + \alpha(1-\alpha) + (1-\alpha)^2\},\\
&p(G|C,F,G) = \alpha^2/\{\alpha^2 + \alpha(1-\alpha) + (1-\alpha)^2\}.
\end{aligned} \tag{3}$$

4. Parameter estimation by ABC

4.1. Fitting values of summary statistics to observed values

The generated biogeographic distribution varied largely among runs of simulation. Instead of fitting the generated distribution by itself to the observed distribution, we fitted the value of summary statistics to the observed value. For each simulation run we calculated the values of summary statistics from the current states on the nearest grids to the sampling locations. These were contrasted with observed values.

To avoid excessive computational costs, the simulation comprised 40 discrete evenly spaced steps. Therefore, the parameter r, the timing of migration, was selected from 39 equally spaced values. Finally, with the function abc of the R (R Core Team, 2017) package abc (Csilléry, François, and Blum, 2012), the posterior distribution of each parameter was obtained using the neuralnet method (Blum and François, 2010) with a tolerance rate 0.025. In total 360,000 runs were conducted to estimate the posterior.

4.2. Two Summary statistics

A) Templeton statistics

Templeton proposed the clade distance (D_c) and the nested clade distance (D_n) in NCPA (Posada et al., 2006; Templeton et al., 1995). The clade distance measures the geographic spread of a clade, and the nested clade distance measures how a clade is geographically distributed relative to other clades (Figure B1). D_c and D_n were calculated as follows. From all the points in the distribution of *N. temminckii*, the centroid point of all the sampling points (C_{All}) and the centroid point of clade X (C_X) were extracted. The centroid point of clade X is the member $i \in cladeX$ that minimizes the average distance to the other members:

$$argmin \sum_{j \in cladeX} d_{ij} \quad (4)$$

where d_{ij} is the geographic distance between points *i* and *j*. Then, D_c and D_n of clade X become:

$$D_c = mean(d_{C_X j})$$

$$D_n = d_{C_X C_{All}} \quad (5)$$

B) Spatial autocorrelation

Spatial autocorrelation (S_a) measures how each clade is aggregated or mixed (Figure B1). S_a is defined as

$$S_a = \sum_{i,j} exp(-cd_{ij}) g_{ij} / \sum_{i,j} exp(-cd_{ij}) \quad (6)$$

Here, g_{ij} is equal to 1 if individuals at points *i* and *j* are members of one clade, while it equals 0 if they belong to different clades. In this study, the value of *c* was set to 0.02. With this value, a point 10 km away has a weight of 0.82 and a point 100 km away has a weight of 0.14.

4.3. Prior distribution

Vague but informative priors of the four parameters were set as: $m \sim unif\{0,5\}, s \sim unif\{0,50\}$, $\alpha \sim unif\{0.5,1\}$, $r \sim unif\{0,1.31\}$, to avoid unduly poor convergence by sampling highly unlikely parameter values at high frequency.

5 Allozyme analysis

Table S3

Genotypes and frequency at the *PEPA* locus and their clade type at the three river populations.

Place	Genotype	Observed number of individuals	Expected number of individuals	Clade type

Kumozu River	*100/*100	0	0.02	-
	*100/*120	1	0.95	C
	*120/*120	12	12.03	-
		X-squared = 0.0208		
Ibi River	*100/*100	1	1.13	F
	*100/*120	7	6.75	F
	*120/*120	10	10.13	C
		X-squared = 0.0247		
Suzuka River	*100/*100	2	1.07	C
	*100/*120	4	5.87	F, F
	*120/*120	9	8.06	-
		X-squared = 1.52		

Allozyme analysis of the *PEPA* locus was conducted as described in Okazaki et al (1991). The upper portions of Kumozu, Ibi and Suzuka rivers located in the western Tokai region, polymorphism was observed caused by *100 and *120 alleles at *PEPA* locus. The observed number of individuals by genotype are consistent with the expected number by the Hardy–Weinberg equilibrium. Several individuals were sequenced from samples. Individuals with genotype *120/*120 were classified into clade C, and genotype *100/*100 into clades C and F. Heterozygous individuals were classified into clades C or F, depending on the sample. The haplotypes of mitochondria and allozyme were not consistent, which means that the individuals over the boundary crossed randomly. We conducted a goodness-of-fit test by simulating the random values from the chi-square distribution (df = 1). We simulated the three random values at each iteration, chose the maximum value, and obtained the distribution of the maximum value. The observed maximum chi-square value was 1.52, which is far lower than the 95th percentile of the simulated distribution (5.65). Observed genotype frequencies are consistent with expected values.

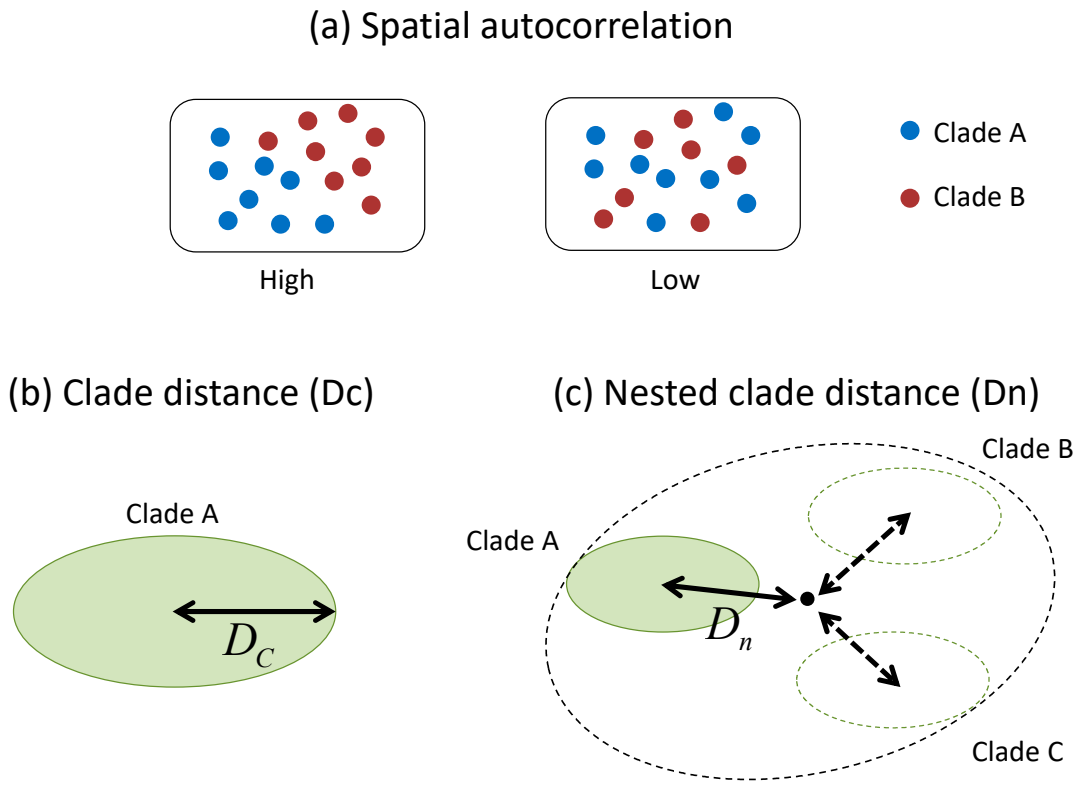

Figure S1. Summary statistics for ABC-based testing hypothesis of intra-species replacement: (a) spatial autocorrelation, measuring how each clade is aggregated or mixed; and (b) clade and (c) nested clade distances, measuring geographic arrangement and expansion of distribution.

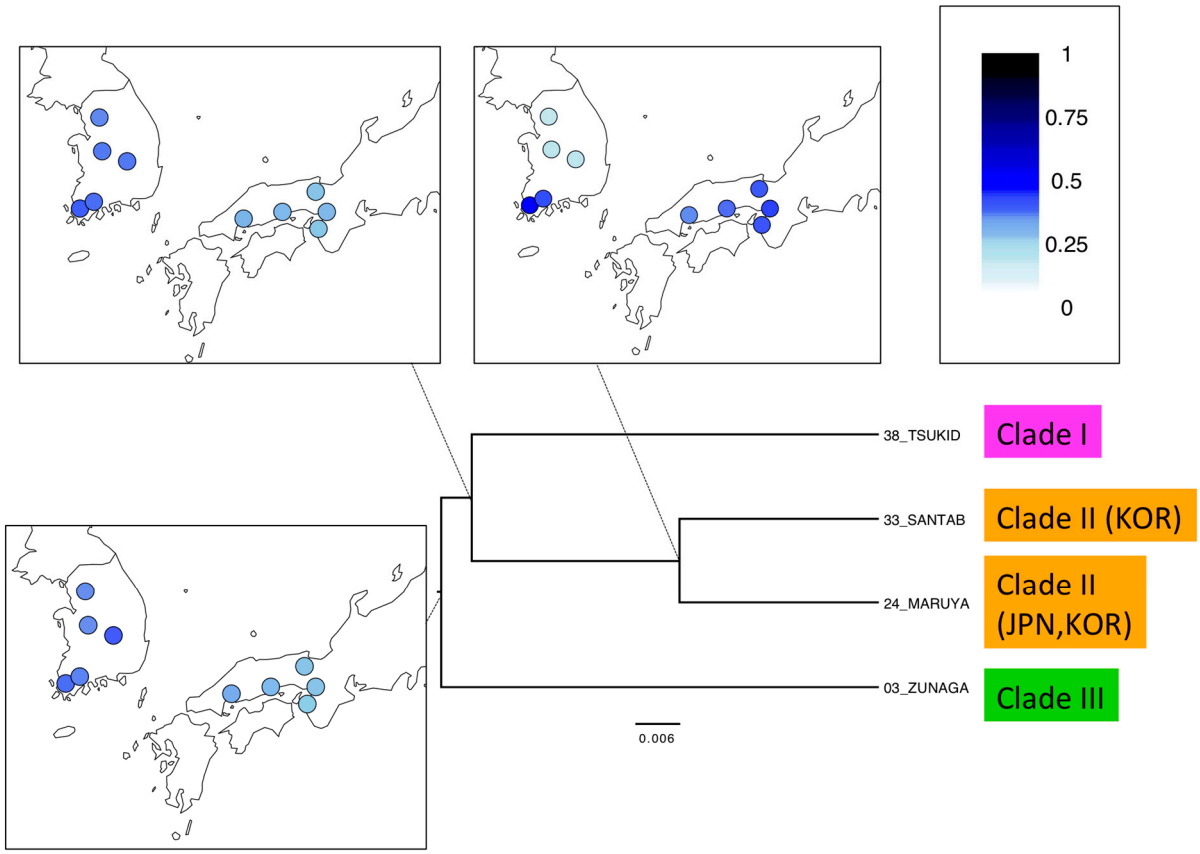

Figure S2. Biogeographic history of *H. longirostris* estimated by BayArea: Ancestral geographic distribution was estimated for each ancestral node. Each circle indicates the discrete area, with the blue coloring representing the posterior probability of existence.

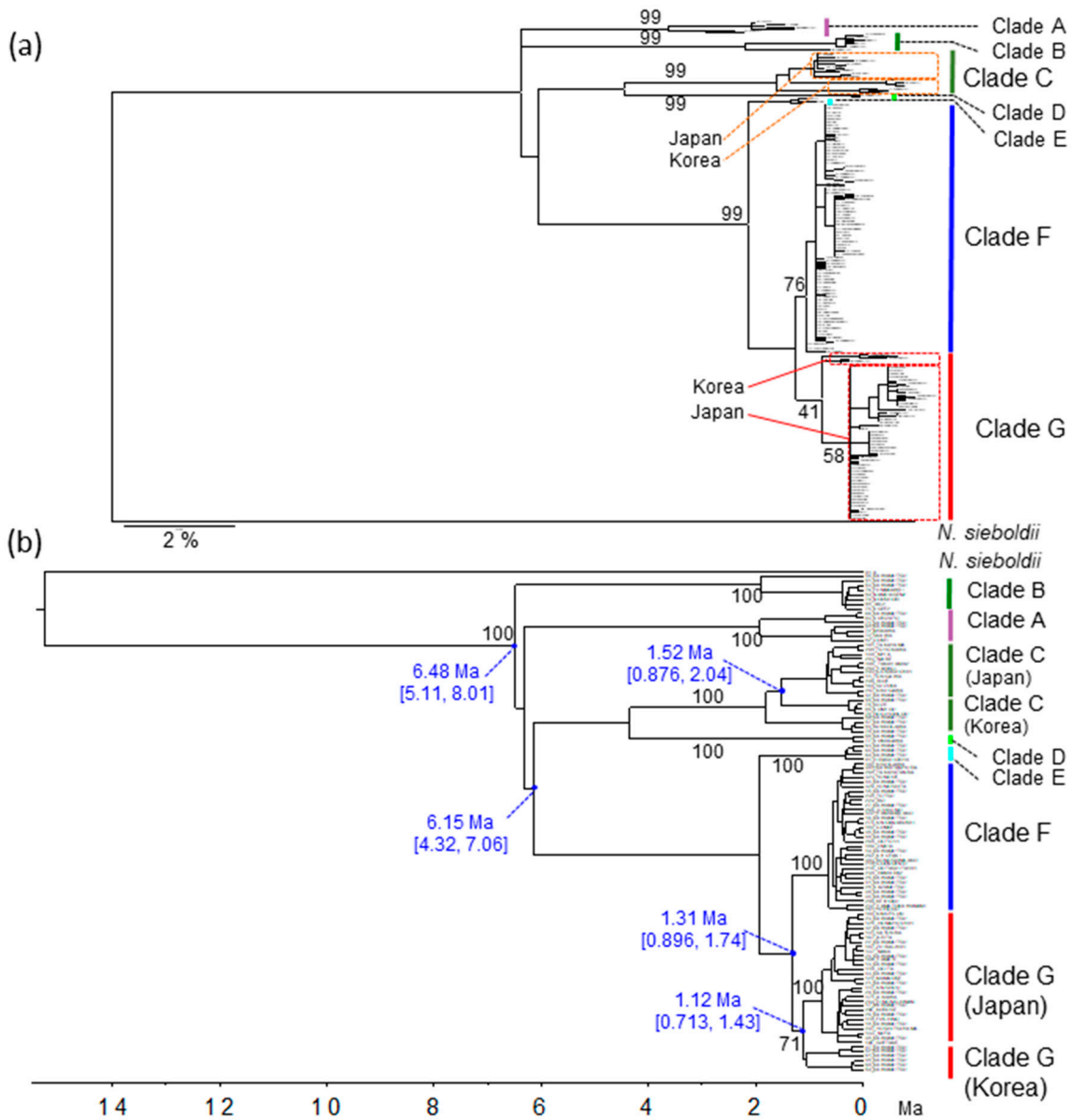

Figure S3. Phylogenetic trees of *N. temminckii* reconstructed from ND II sequences. (a) ML tree, (b) Bayesian tree. Black numbers in phylogenetic trees indicate bootstrap or posterior probabilities (%). Blue numbers without brackets in (b) indicate point estimates of node age; blue numbers within brackets indicate 95% upper and lower credibility interval limits (Ma). The topology of the Bayesian tree is slightly different from that of the ML tree and Figure 4a. This difference is partly because of the different estimation procedure. In any case, the sequence of *N. sieboldii* was added in the estimation of the divergence times, whereas, in the ancestral state reconstruction of the distributed area, the outgroup was not included. These inconsistencies in topology were confined to the root region of the tree and do not affect the simulation scenario and resulting conclusion.

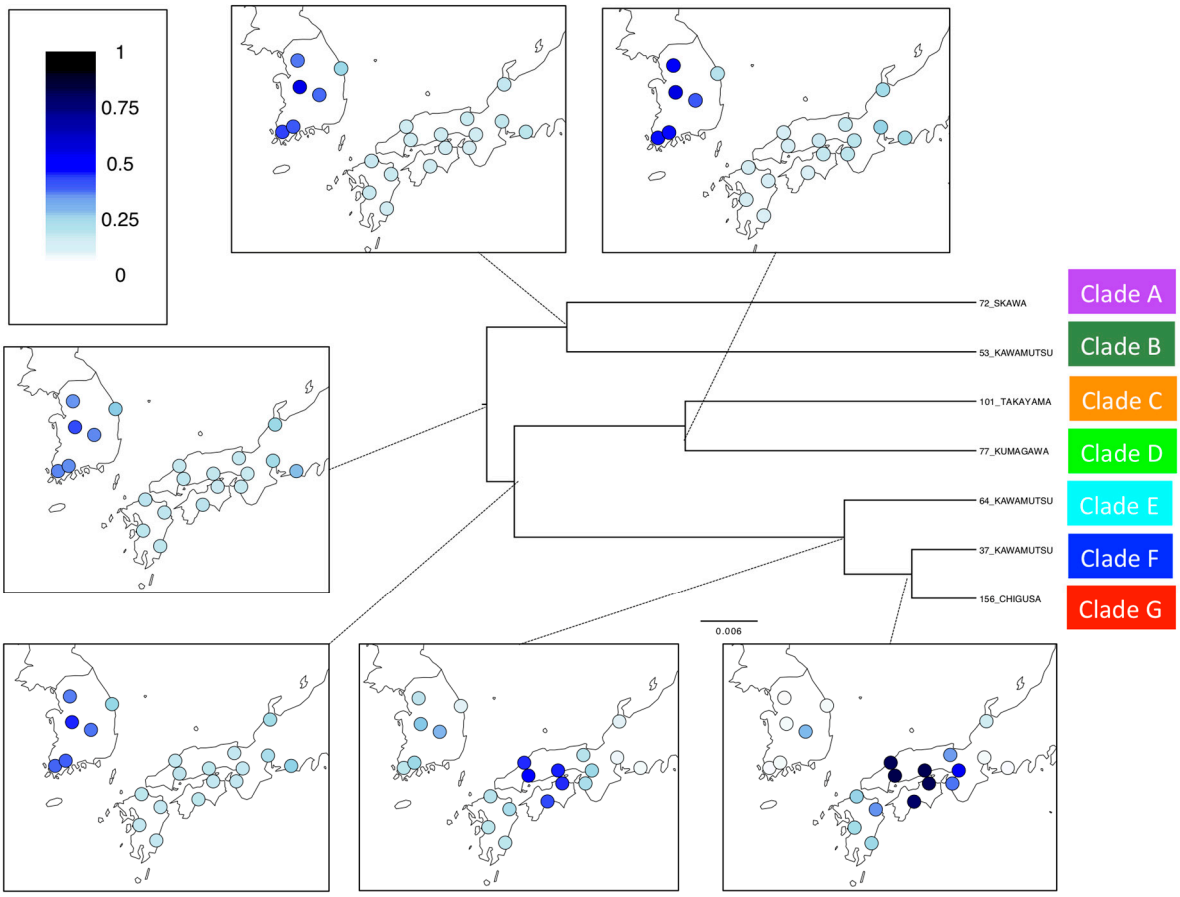

Figure S4. Biogeographic history of *N. temminckii* estimated by BayArea: Ancestral geographic distribution was estimated for each ancestral node. Each circle indicates the discrete area, with the blue coloring representing the posterior probability of existence.

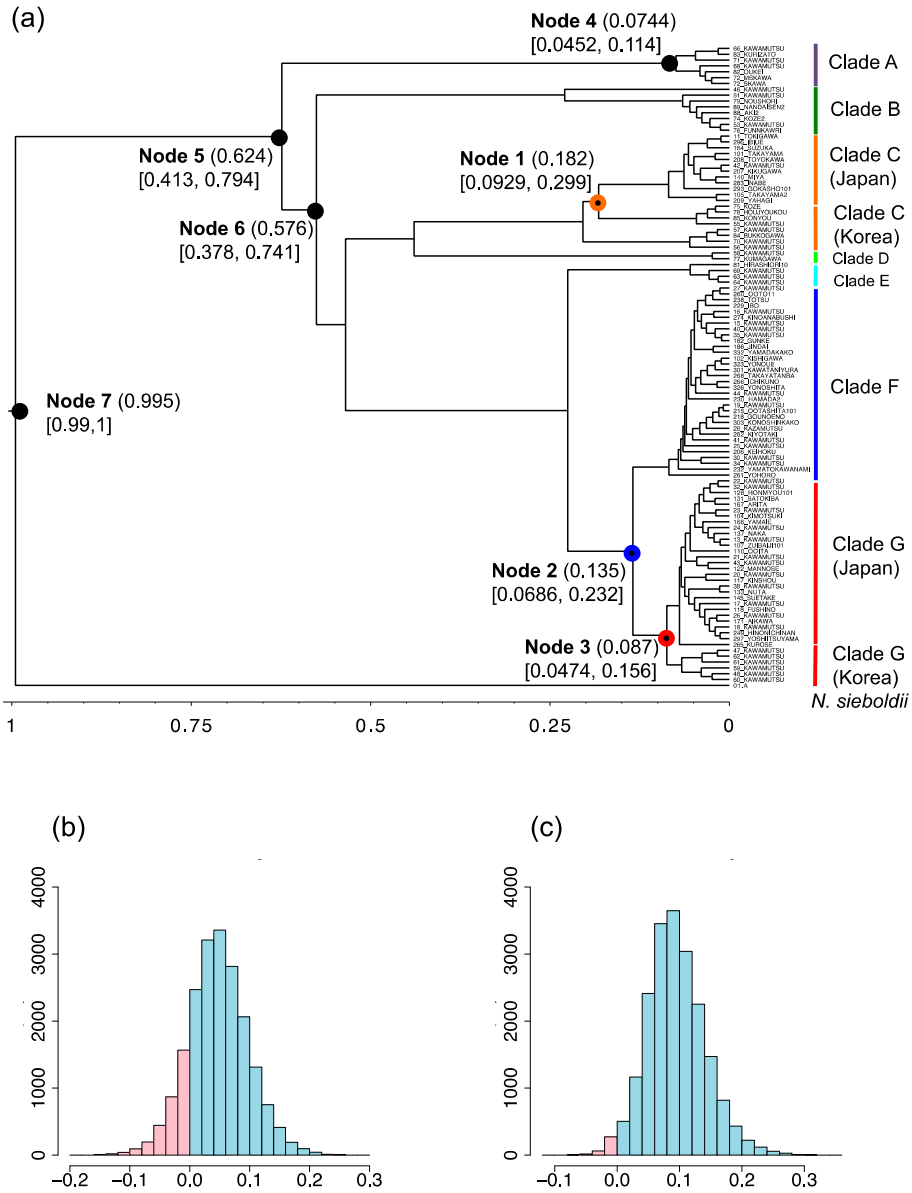

Figure S5. Divergence time estimation by MCMCTREE: (a) Estimated relative divergence time of nodes in parentheses beside node number, and 95% credibility interval below the node number; (b) and (c) histograms of MCMC samples. (b) Difference in relative divergence time between nodes 1 and 2; the probability of node 1 being older than node 2 is 83.8% in blue, in contrast to the opposite event in red. (c) Difference in relative divergence time between nodes 1 and 3; the probability of node 1 being older than node 3 is 98.2%.

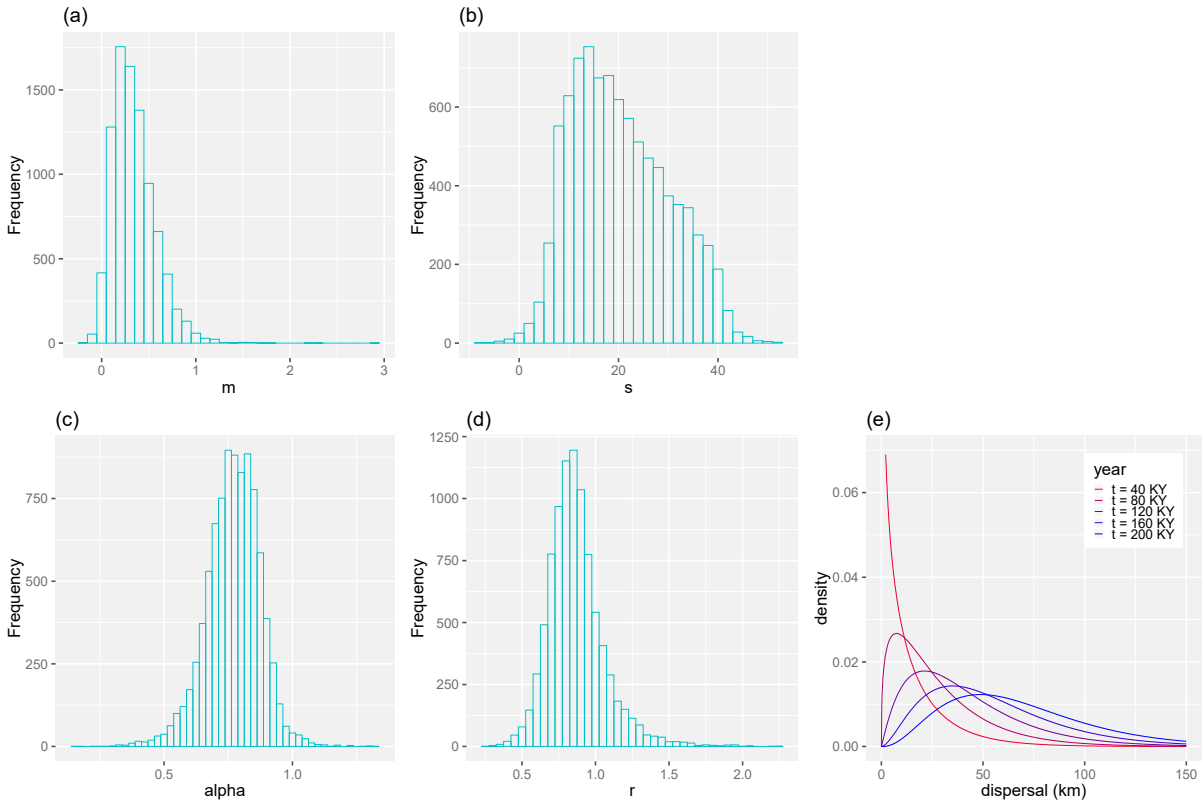

Figure S6. Posterior distribution of four parameters, and probability density of dispersal distance: Posterior distributions of (a) m, (b) s, (c) α, and (d) r. (e) The probability distribution of dispersal distance under the point estimates of parameter m (0.345) and s (20.2).

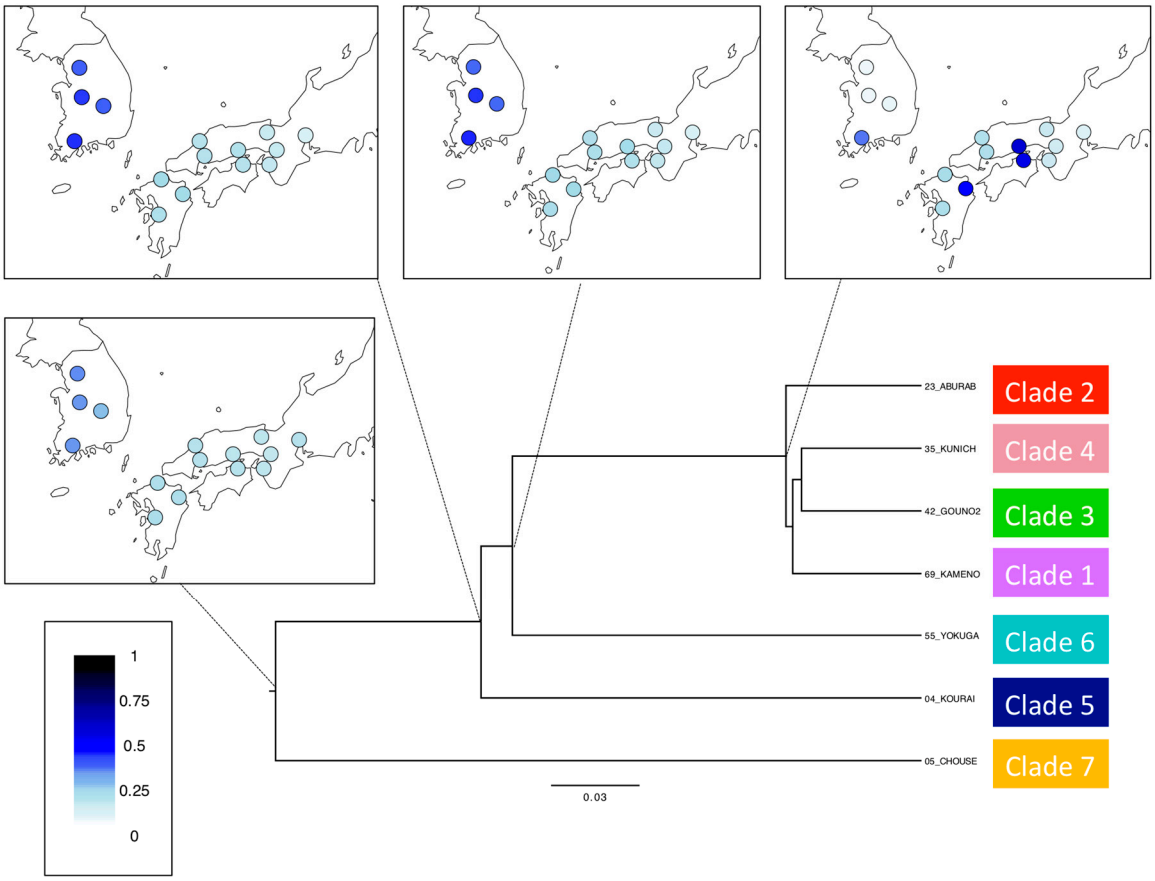

Figure S7. Biogeographic history of *T. limbata* and related species estimated by BayArea: Ancestral geographic distribution was estimated for each ancestral node. Each circle indicates the discrete area, with the blue coloring representing the posterior probability of existence.

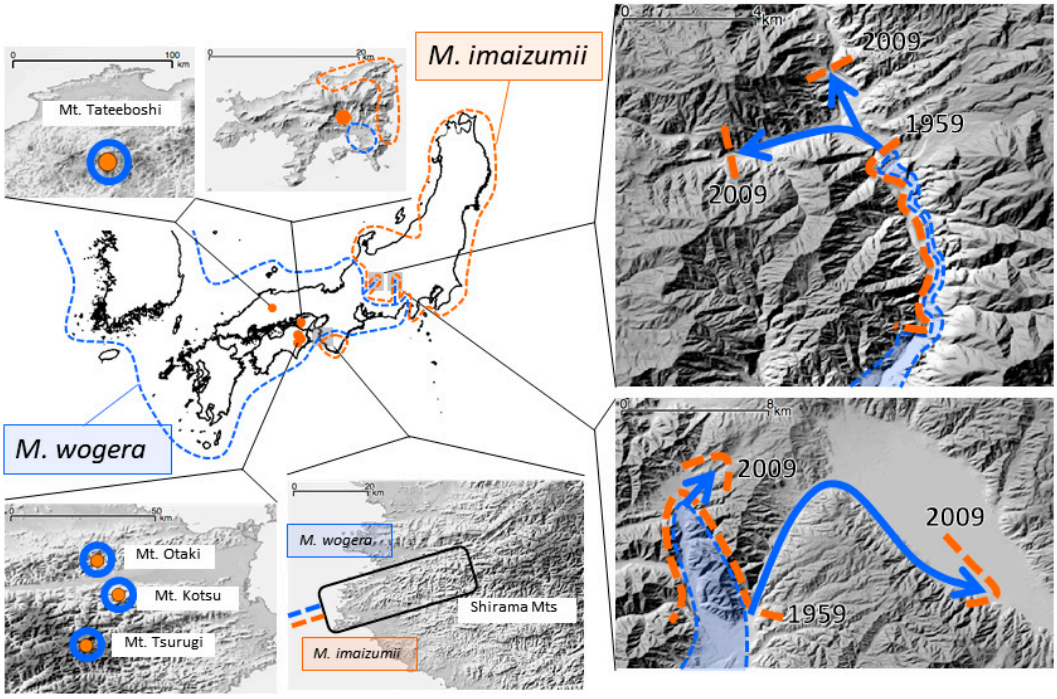

Figure S8. Distribution of *Mogera wogura and M. imaizumii*. At the eastern edge of the distribution of *M. wogura*, this species has reputedly expanded its distribution and replaced *M. imaizumii* between 1959 and 2009. In the Chugoku and Shikoku regions, the distribution of *M. imaizumii* is isolated to a habitat atop several mountains or an island within the Seto Inland Sea; this species is also distributed in the southern Kinki region, with Shirama Mountains at the boundary. Map based on descriptions by Abe (1995, 2001, 2010). The original elevation chart (in color) was provided by the Geospatial Information Authority of Japan; marine areas were assembled using data from the Hydrographic and Oceanographic Department, Japan Coast Guard (Geospatial Information Authority of Japan, 2013).

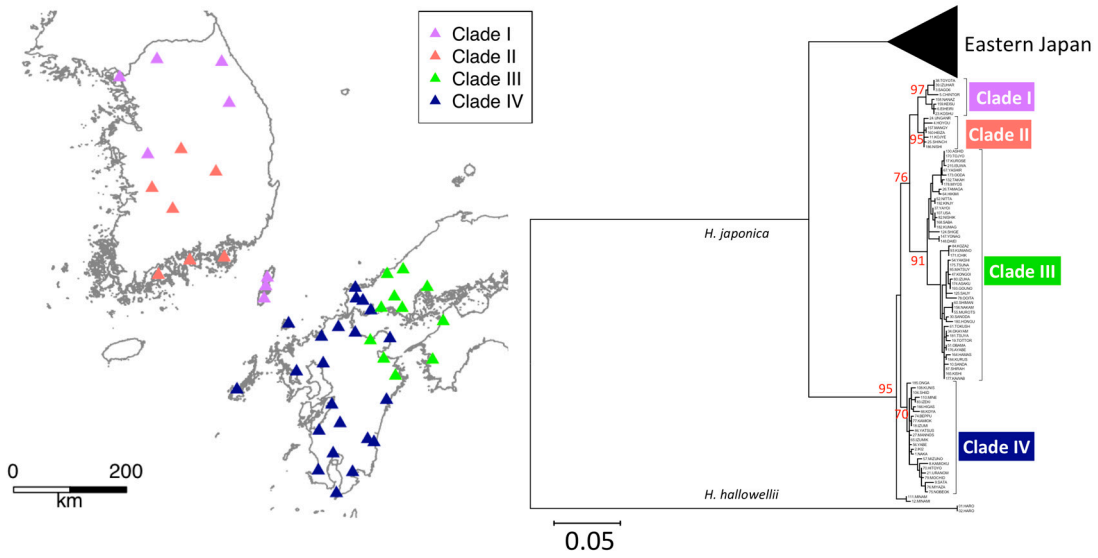

Figure S9. Phylogenetic trees of ND II sequences and geographic locations: *Hyla japonica*. ML phylogenetic trees and biogeographic maps for *H. japonica* were obtained in a similar fashion to those for *N. temminckii*. The red numbers were the bootstrap probabilities. We used sequences of *H. hallowellii* as an outgroup for *H. japonica*. Clade I was sampled in the middle of the Korean Peninsula and on Tsushima Island, Japan, while clade II was sampled in southern Korea. For *H. japonica*, the Korean Peninsula provided a suitable habitat even during the last glacial maximum (Dufresnes et al., 2016), so it is unlikely that extinctions were caused by climate change. The distribution of clade I was divided by that of clade II. The average distance between *H. japonica* clades I and II is 1.87%, at the intra-species level. The obtained sequences were deposited in DDBJ/ENA/GenBank (accession numbers were LC568290–LC568534).

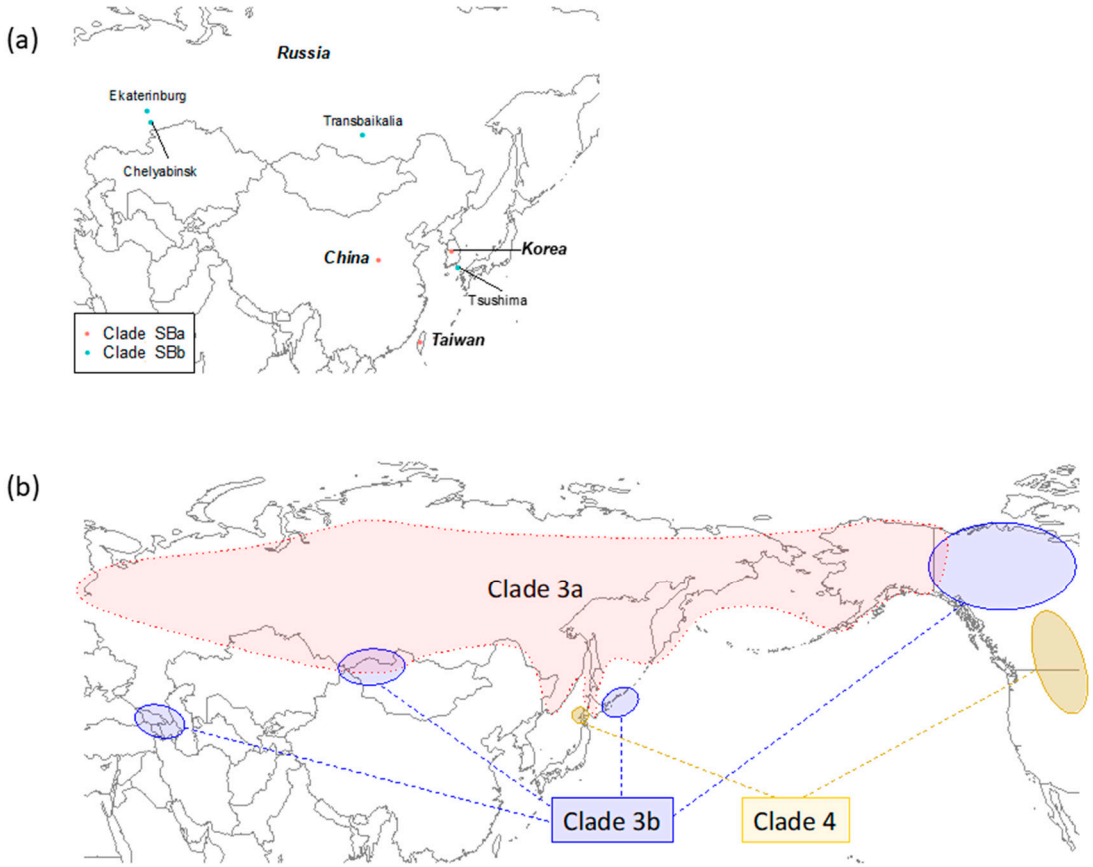

Figure S10. Distribution maps of Siberian weasel (a) and Brown bear (b): (a) Modified from Shalabi et al. (2017). Clade SBa was sampled in China, Korea and Taiwan, while clade SBb was mostly sampled in Russia, but also Tsushima Island. The distribution of clade SBb is divided into Russia and Tsushima Island, while the distribution of clade SBa exists in between. (b) Based on previous studies (Hirata et al., 2013; Hirata et al., 2014; Waits et al., 1998) the distribution of clade 3a continuously expands from Russia to Alaska, while the distributions of other clades are divided. Clade 3b has several isolated populations on the Asian continent, around Japanese Hokkaido, and in North America. These isolated populations exist near the periphery of the continuous distribution of clade 3a. The distribution of clade 4 is also divided. Of the three clades, clade 4 in North America is farthest from Bering Strait; a population in the western area of Japanese Hokkaido also exists.

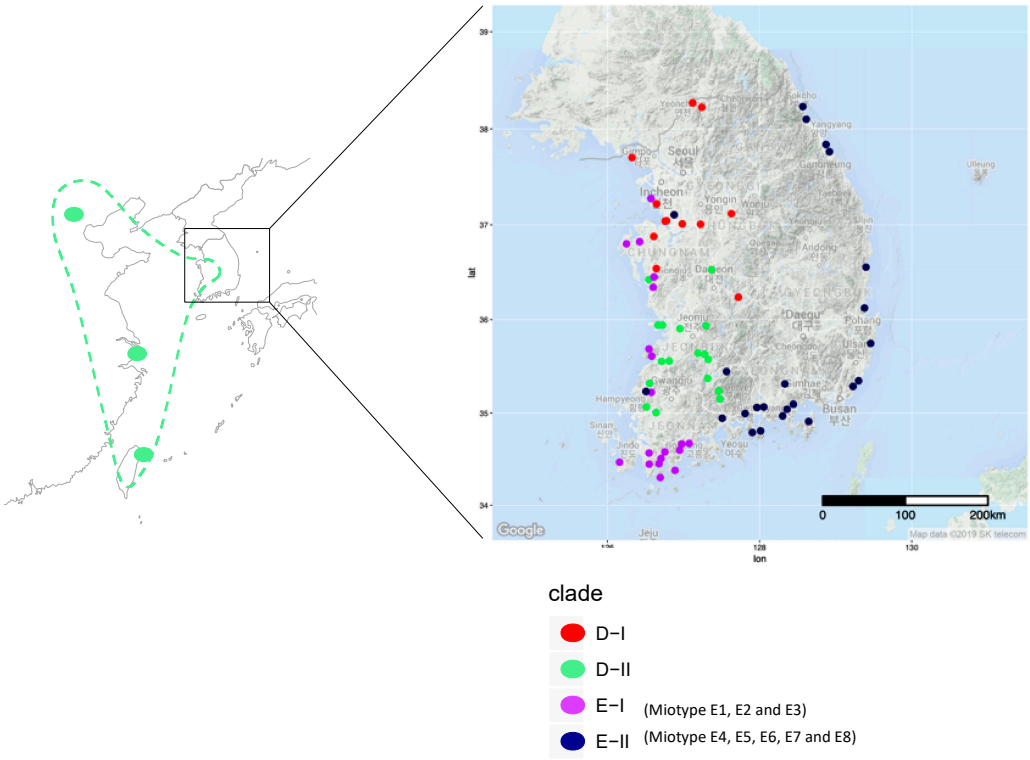

Figure S11. Distribution maps for *Oryzias sinensis*, based on Takehana et al. (2004). Two clades (D, E) and two subclades of D (D-I and D-II) of *O. sinensis* are recognized; we also split clade E into subclades E-I and E-II. Subclade D-II has a distribution from China to southwestern Korea, and clade D-I inhabits midwestern Korea. Clade E-I occurs in the most southwestern part of Korea, with fragmented distributions where clade D-I and D-II are widely distributed. Clade E-II has a wide distribution in Eastern Korea, and a fragmented distribution near Incheon and Hampyeong. Map drawn using the R (R Core Team, 2017) package ggmap (Kahle and Wickham, 2013).

References

Aoyama, J., Watanabe, S., Ishikawa, S., Nishida, M., & Tsukamoto, K. (2000). Are morphological characters distinctive enough to discriminate between two species of freshwater eels, *Anguilla celebesensis* and *A. interioris*? *Ichthyological Research*, *47*, 157–161.

Asahida, T., Kobayashi, T., Taitoh, K., & Nakayama, I. (1996). Tissue preservation and total DNA extraction from fish stored at ambient temperature using buffers containing high concentration of urea. *Fisheries Science*, *62*, 727–730.

Blum, M. G. B., & François, O. (2010). Non-linear regression models for Approximate Bayesian Computation. *Statistics and Computing*, *20*, 63–73. https://doi.org/10.1007/s11222-009-9116-0

Chang, Y., Huang, F., & Lo, T. (1994). The complete nucleotide sequence and gene organization of carp (*Cyprinus carpio*) mitochondrial genome. *Journal of Molecular Evolution*, *38*, 138–155.

Dufresnes, C., Litvinchuk, S. N., Borzée, A., Jang, Y., Li, J. T., Miura, I., … Stöck, M. (2016). Phylogeography reveals an ancient cryptic radiation in East-Asian tree frogs (*Hyla japonica* group) and complex relationships between continental and island lineages. *BMC Evolutionary Biology*, *16*(1), 1–14. https://doi.org/10.1186/s12862-016-0814-x

Hall, H. J. & Nawrocki, L. W. (1995). A rapid method for detecting mitochondrial DNA variation in the brown trout, *Salmo trutta*. *Journal of Fish Biology, 46,* 360–364.

Higgins, D. G., & Sharp, P. M. (1988). CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. Gene, 73(1), 237-244.

Kahle, D., & Wickham, H. (2013). ggmap: Spatial Visualization with ggplot2. *The R journal*, *5*(1), 144–161.

Karney, C. F. F. (2013). Algorithms for geodesics. *Journal of Geodesy*, *87*, 43–55. https://doi.org/10.1007/s00190-012-0578-z

Palumbi, S., Martin, A., Romano, S., McMillan, W. O., Stice, L., & Grabowski, G. (1991). *Simple fool's guide to PCR*. Honolulu, HI: University of Hawaii.

Waits, L., Talbot, S., Ward, R. H., & Shields, G. (1998). Mitochondrial DNA phylogeography of the North American brown bear and implications for conservation. *Conservation Biology*, *12*, 408–417.