Title:

Medaka population genome structure and demographic history unveiled via Genotyping-by-Sequencing

Authors and Affiliations:

Takafumi Katsumura[1,2*], Shoji Oda[3], Mitani Hiroshi[3], Hiroki Oota[1*].

[1] Department of Anatomy, Kitasato University School of Medicine
1-15-1 Kitasato, Minami-Ku Sagamihara, Kanagawa 252-0374, Japan

[2] Graduate School of Natural Science and Technology, Okayama University, 1-1-1 Tsushimanaka, Kita-Ku, Okayama 700-8530, Japan

[3] Department of Integrated Biosciences, Graduate School of Frontier Sciences, University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa, Chiba 277-8562, Japan

***Authors for Correspondence:**

Takafumi Katsumura, Graduate School of Natural Science and Technology, Okayama University, Okayama, Japan, +81 86 252 7860, tk@okayama-u.ac.jp

Hiroki Oota, Department of Anatomy, Kitasato University School of Medicine, Kanagawa, Japan, +81 42 778 9022, hiroki_oota@med.kitasato-u.ac.jp

Abstract

Medaka is a model organism in medicine, genetics and developmental biology research and in population genetics studies, and lab stocks composed of more than 100 local wild populations are available for such research. Thus, medaka represents a potentially excellent bioresource for screening disease-risk- and adaptation-related genes by a genome-wide association study (GWAS). Although a genetic population structure should be known beforehand with practicing a genome-wide association study (GWAS), a comprehensive study on the genome-wide diversity of wild medaka populations has not been performed. Here, we performed Genotyping-by-Sequencing (GBS) for 81 and 12 medakas captured from a bioresource and the wild, respectively. Based on the GBS data, we evaluated the genetic population structure and estimated the demographic parameters using an approximate Bayesian computation (ABC) framework. The autosomal data confirmed high differentiation among local populations and supported our previously proposed hypothesis on medaka dispersal that was based on mitochondrial genome (mtDNA) data. A new finding was that a local group thought to be a hybrid between the northern and the southern Japanese groups was actually a sister group of the northern Japanese group. This paper presents the first population-genomic study of medakas and provides information indispensable for a GWAS of this fish.

Key words: local population, freshwater fish, demography, RAD-seq, bioresource

Introduction

A genome-wide association study (GWAS) is a powerful method of detecting causal genes for various phenotypic traits (Welter et al. 2013). However, molecular engineering performed to reveal the functional differences in the genetic variation observed in GWASs represents a technical and ethical bottleneck for humans. Instead of humans, inbred/outbred animals and plants derived from wild sources have often been used as model populations for GWASs (Mackay et al. 2012; Kirby et al. 2010; Gonzales & Palmer 2014; Parker et al. 2016). Wild strains should provide more persuasive data for detecting natural traces related to adaptation than inbred/outbred strains maintained in research institutes because wild populations may represent potentially effective resources for identifying genes on which natural selection has acted within the ecological environment. Current molecular biological techniques can be applied to wild species of model organisms, and they can reveal the function of candidate genes detected by a GWAS. Nevertheless, a GWAS using such wild strains has been not conducted, and the only results have been obtained for limited species, such as *Arabidopsis thaliana* (Consortium et al. 2016). In general, usable and accessible bioresources for a species, which harbors within-species genetic diversity (in particular, genome population structure), are difficult to maintain.

Medaka (*Oryzias latipes*) is a small fresh-water fish native to East Asia, and it is a rare vertebrate model with the potential for use as a model organism for a GWAS using wild strains. Medaka wild strains have been maintained in universities and research institutes as a bioresource (hereafter, wild lab stocks) that consist of more than 100 local-wild populations showing various phenotypic traits (Watanabe-Asaka et al. 2014; Igarashi et al. 2017) and enormous genetic diversity. Medakas can be bred at a lower cost and mature in shorter timeframes (approximately 2 months) than mice. Our ultimate goal in exploiting medaka characteristics is to establish an experimental system for testing the functional differences

65 between the alleles detected using a GWAS (Oota & Mitani 2010). Once genetic
66 polymorphisms related to phenotypic traits in medaka are detected, *in vivo* experiments, such
67 as genome editing experiments (Ansai & Kinoshita 2014), must be conducted to understand
68 the variant function; however, such experiments are ethically unacceptable in humans.
69 Revealing the functional difference between the alleles in wild populations allows us to infer
70 the role of genetic polymorphisms in human homologous genes. If the phenotypic traits are
71 significantly associated with adaptations to the medaka population's local environment, we
72 may be able to speculate about each allele's role in polymorphic sites in the context of human
73 adaptation because medakas are vertebrates that are genetically closer to humans than fruit
74 flies and Arabidopsis.
75 However, few genome-wide studies have been conducted on the medaka population
76 structure. Most previous studies on medaka populations have been conducted using
77 mitochondrial DNA (mtDNA). Medaka is divided into four mitochondrial groups: the
78 northern Japanese (N.JPN), southern Japanese (S.JPN), eastern Korean (E.KOR) and western
79 Korean/Chinese (W.KOR) groups (Sakaizumi et al. 1980; 1983; Sakaizumi 1986; Katsumura
80 et al. 2009). An ambiguous group occurs in the geographic boundary region known as
81 Tajima-Tango (see also fig. 1) between N.JPN and S.JPN and is thought to be a hybridization
82 of the two groups, as defined by the mtDNA data (Matsuda et al. 1997). N.JPN and S.JPN
83 are composed of 3 and 11 subgroups, respectively (Takehana et al. 2003). Each subgroup is
84 composed of local populations, and between-population genetic diversities are greater than
85 within-population genetic diversities, indicating high genetic differentiation between local
86 populations (Katsumura et al. 2009). Although the origin of S.JPN with the largest habitat
87 has been suggested as the northern part of Kyushu Island, we have proposed the "Out of
88 Northern Kyushu" hypothesis instead (Katsumura et al. 2012). However, mtDNA is
89 insufficient to describe the population structure and history because it is a single locus.

90 Although genetic diversity based on autosomes should be indispensable to the medaka GWAS,
91 it remains extremely limited.
92 The lack of autosomal information increases the difficulty of recognizing the cryptic
93 population structure (i.e., the population structure inferred from the geographical distribution
94 is not necessarily consistent with the genetic information) because the possibility of a false
95 positive/negative caused by unintended sampling bias is increased (Brachi et al. 2011). In
96 this study, we comprehensively performed a population genetic analysis based on autosomal
97 single nucleotide polymorphisms (SNPs) using maintained wild lab stocks that include 81
98 local populations. We sampled one individual from each population to estimate the
99 population structure and history. Additionally, we captured 12 wild individuals from the
100 northern part of Kyushu Island, where the medakas currently distributed throughout
101 Pacific-side Japan originated (Katsumura et al. 2012), to estimate the time of medaka
102 dispersal Out of Northern Kyushu. To obtain SNP data, we conducted
103 Genotyping-by-Sequencing (GBS) (Elshire et al. 2011; Narum et al. 2013) using a
104 high-throughput sequencer. This method, which allows us to cost-effectively gain more than
105 tens of thousands of SNPs (Andrews et al. 2016), results in an accurate enhancement of the
106 population genetic estimations because the variance in the demographic parameter estimates
107 are decreased when we used many SNPs on the genome. Eventually, we obtained more than
108 ten thousand SNPs from eighty-one wild lab stocks and twelve wild-captured medakas.
109 Here, we re-evaluate medaka genetic diversity and population structures based on these SNPs
110 and reconstruct the population history by assessing three demographic events. These data
111 allow us to conduct the appropriate sampling needed to conduct a successful GWAS and
112 provide a basis for discussing the role of phenotype-associated alleles in the context of
113 adaptation.

Materials and Methods

Samples

We sampled 81 male medakas from 81 wild lab stocks (14 from N.JPN, 56 from S.JPN, and 5 from Tajima-Tango in the geographic boundary region between N.JPN and S.JPN, 3 from E.KOR, and 3 from W.KOR) at the University of Tokyo in 2014. The lab stocks originated from geographically distinct populations in East Asia and have been maintained since 1985 (Shima et al. 1985) as closed colonies in the Graduate School of Frontier Sciences, the University of Tokyo, Kashiwa City (fig. 1 and supplementary fig. 1). These lab stocks maintained the genetic diversity originating from their habitat (Katsumura et al. 2014) and show less diversity within populations than between populations (Katsumura et al. 2009). Because of these characteristics, we considered that one individual sampled from the lab stock would be adequate to represent its originating population. In addition, we used two wild-captured medaka populations from the Saga Prefecture in the northern part of Kyushu, Japan. One population is the Ogi (SO) population from Southern Saga, and the other population is the Umeshiro (US) population from Northern Saga, both of which were captured in September 2010 (see Katsumura *et al*. 2012). Six of 48 medakas from each wild-captured population were selected randomly and analyzed via GBS.

DNA extractions and Genotyping-by-Sequencing

One-third of the medaka body was dissolved in a 600 µl lysis buffer containing 1.24% SDS, 0.124 M EDTA and 0.062 mg/ml proteinase K (final concentrations). The total genomic DNA was extracted and purified using phenol-chloroform and isopropanol precipitation. After 70% EtOH washing, an isolated DNA pellet was resuspended in a 100 µl TE buffer and then treated with RNase A (final conc. 1 mg/ml) for 1 hour at room temperature; then, the DNA was purified again using phenol-chloroform and isopropanol precipitation again. For

139 ninety-three samples, the GBS process was outsourced to Macrogen Japan in Kyoto. The
140 procedures for constructing libraries and performing Illumina HiSeq 2000 single-end
141 sequencing were the same as those described by Poland *et al.* (Poland et al. 2012) except for
142 the use of the restriction enzyme ApeKI instead of EcoRI–MspI. The sequence lengths were
143 51 bp and included each individual in-line barcode (4–9 bp) for the individual sample. The
144 data have been submitted to the DDBJ Seaquence Read Archive (DRA) database under
145 project accession ID: DRA006353.
146
147 **Quality filtering and SNP extraction**
148 Our single-end reads were filtered using *FASTQ Quality Filter* in *FASTX-Toolkit version*
149 *0.0.13* (http://hannonlab.cshl.edu/fastx-toolkit) using the following options:
150 -Q 33 -v -z -q 30 -p 90. The draft genome of the medaka sequenced by the PacBio
151 sequencer (Medaka-Hd-rR-pacbio_version2.0.fasta; unpublished data from
152 http://utgenome.org/medaka_v2/#!Assembly.md) was used to align the reads using
153 *BWA backtrack 0.7.12-r1039* (H Li & Durbin 2009) using the "-n 0.06" option. After the
154 mapping process, the multi-mapped reads were removed using *Samtools v1.2* (H Li et al.
155 2009) and the "-Sq 20" option. Following this pipeline, sequencing of the ApeKI-digested
156 GBS libraries generated an average of 3.06 million reads per individual before any quality
157 filtering. The read numbers ranged from 1.89 to 4.10 million reads per individual. After
158 quality filtering, 2.76 million (90.2%) sequences per individual on average were retained and
159 0.30 million (9.8%) sequences were eliminated. The retained sequences presented a mean
160 quality score of 38.7 and a GC content of 47.8%. An average of 1.84 million of the retained
161 sequences (66.7%) aligned to the medaka autosomal genome, and 0.92 million sequences
162 (33.3%) were not aligned and discarded because they mapped to non-autosomal
163 (mitochondrial genome and unanchored contigs) and multiple loci.

The *Stacks* pipeline (version 1.35) and the Stacks workflow (https://github.com/enormandeau/stacks_workflow) were used to generate SNPs and sequences for each individual separately (Catchen et al. 2011; Catchen, Hohenlohe, et al. 2013). The selected *Stacks* parameters were as follows: minimum stack depth (-m), 3; and number of mismatches when building the catalog (-n), 1. In the 'rx' step of *Stacks*, we used the bounded SNP model set to 0.1, the ε upper bound set to 0.1 and the log likelihood set to 10. Following this pipeline, five datasets were constructed using the *population* program in *Stacks*.

The first is the "PopStat" dataset, which was generated using the "-p 5 -r 0.66" options and geographical grouping (N.JPN, Tajima-Tango, S.JPN, E.KOR, W.KOR; see also fig. 1) to calculate the population genetic statistics (table 1) using the loci shared across all groups and sequenced in one or more populations in each group.

The second is the "Global" dataset, which was generated using the "-p 58 -r 1.00" option and without the geographical grouping to examine a phylogenetic relationship between the geographic populations and population structure within the species.

The third and the fourth are the "HZ-1" and "HZ-2" datasets, respectively, including the 15 boundary populations (N.JPN: Kaga, Maiduru, Miyadu, Obama, and Sabae; Tajima-Tango: Amino, Hamasaka, Kinosaki, Kumihama, and Toyooka; and Honshu: Ayabe, Iwami, Kasumi, Matsue, and Tottori) with or without the Kyushu populations (Kyushu: Fukue, Hiwaki, Izumi, Kadusa, and Kikai). These datasets were generated using the "-p 3 -r 0.70" (without Kyushu) or "-p 4 -r 0.70" (with Kyushu) options, respectively, to assess the genetic population structure and the history of the boundary population for the Tajima-Tango group.

The fifth is the "Local" dataset, including two Kyushu deme samples: one is Umeshiro (US), which has been sampled in the northern part of the Saga prefecture, and another is Ogi (SO), which has been sampled in the southern part of the Saga prefecture.

189 These samples were used to estimate the time of Honshu's population divergence from
190 Kyushu to infer the timing of the "Out of Northern Kyushu" event (Katsumura et al. 2009).
191 Note that we define the term "deme samples" and "non-deme samples" as "samples
192 from the local-wild population" and "samples from the wild lab stocks", respectively (see
193 details in Katsumura *et al*. 2009).
194 Against those datasets (except "PopStat" and "HZ-1"), the SNPs with strong linkage
195 disequilibrium were randomly removed, one from each pair of SNPs with $r^2 > 0.2$, using
196 Plink1.9 (Chang et al. 2015) and the "--indep-pairwise 12.5 5 0.2 --autosome-num 24" option.
197 These values were set based on a medaka population genomics study (Spivakov et al. 2014).
198 Finally, "Global", "HZ-2" and "Local" datasets contained 8,361 SNPs out of 13,177 SNPs,
199 1,014 SNPs out of 2,661 SNPs and 698 SNPs out of 2,246 SNPs, respectively.
200
201 **Genetic clustering analysis**
202 To obtain a genetic overview of the relationship of medaka geographic populations, we
203 performed a principal component analysis (PCA) of the "Global" dataset as implemented in
204 the *SNPRelate* program (Zheng et al. 2012) in R version 3.2.2. Additionally, to examine the
205 genetic relationship among medakas in the Japanese archipelago, we used the subdataset
206 without the E/W.KOR and Chinese populations, which included 7,126 SNPs. We also
207 performed a model-based genetic clustering analysis using *ADMIXTURE v1.23* (Alexander et
208 al. 2009) to estimate the proportions of ancestral medaka populations. We ran 50 replicates
209 with random seeds for the number of clusters (K) from 1 to 9 and calculated the mean of the
210 lowest fivefold cross-validation errors for each K (Jeong et al. 2016). K = 4, 5, and 6 show
211 the first-, second- and third-lowest fivefold cross-validation errors, respectively
212 (supplementary fig. 2). The results of the genetic clustering analyses were visualized by a
213 ggplot2 package in R (Wickham 2011).

214

215 **Reconstructing the phylogenetic tree using the maximum likelihood method**

216 We chose an individual as representative of a population and generated the individual
217 sequences using the *population* program with the "-p 58 -r 1.00 --phylip_all" option in the
218 *Stacks* pipeline. The dataset included 4,638 partitions and 217,986 bp nucleotide sequences
219 to compensate for the loci that could not be sequenced by invariable sites across all samples
220 and "N" at the variable site. The dataset including 4,638 partitions and 217,986 bp
221 nucleotide sequences was analyzed via the *IQ-TREE* program (Nguyen et al. 2014) to
222 reconstruct a maximum likelihood tree with model selection for each partition (Chernomor et
223 al. 2016) and 1,000 SH-aLRT/ultrafast-bootstrappings (Guindon et al. 2010; Minh et al. 2013).
224 Then, we used the *FigTree* program (http://tree.bio.ed.ac.uk/software/figtree/) to visualize the
225 phylogenetic tree.

226

227 **Inference of demographic parameters of the populations in and around Tajima-Tango**
228 **and the time of Out of Northern Kyushu based on the ABC framework**

229 Using the HZ-2 dataset, including 15 boundary and 5 Kyushu populations, the population
230 history, including the possibility that the Tajima-Tango group occurred by the admixture of
231 N.JPN and S.JPN, was inferred from 1,014 SNPs using *DIYABC ver2.1.0* (Cornuet, Pudlo,
232 Veyssier, Dehne-Garcia, Gautier, Leblois, Marin & Estoup 2014b). We tested four
233 evolutionary scenarios in which N.JPN and S.JPN diverged from an ancestral population at
234 time t3 (fig. 5a): (I) Tajima-Tango originated in N.JPN: Honshu and Kyushu diverged at time
235 t2 and then Tajima-Tango diverged from N.JPN at time t1 (Takehana et al. 2016); (II)
236 Admixture of N.JPN and Honshu: Honshu and Kyushu diverged at time t2 and then
237 Tajima-Tango occurred at time t1 by an admixture with rate r between N.JPN and Honshu
238 (Sakaizumi 1984); (III) N.JPN originated in Tajima-Tango: Honshu and Kyushu diverged at

239 time t2 and then N.JPN diverged from Tajima-Tango at time t1; and (IV) Honshu diverged
240 from Kyushu and then Tajima-Tango occurred by an admixture with Honshu of rate r. For
241 simplicity, all populations are assumed to have constant effective sizes in each lineage (i.e., no
242 bottleneck and expansion). The same prior parameters were defined for four scenarios based
243 on previous studies, and the prior distribution of each parameter is presented in table 2. In
244 addition, we set the conditional constraint as follows: $t3 > t2$, $t3 > t1$ and $t2 \geq t1$ (see also fig.
245 5a). In total, four million simulations were run, which provided approximately one million
246 simulations for each scenario.

247 The summary statistics for all SNP loci included the proportion of loci with null
248 gene diversity, mean gene diversity across polymorphic loci, proportion of loci with null Nei's
249 distance between the two samples, variance across loci of non-null Nei's distances between
250 two samples, proportion of loci with null admixture estimates, mean across loci of non-null
251 admixture estimates, variance across loci of non-null admixture estimates and mean across all
252 locus admixture estimates. The 1% simulated datasets closest to the observed dataset were
253 used to estimate the posterior parameter distributions through a weighted local linear
254 regression procedure (Beaumont et al. 2002). Scenarios were compared by estimating their
255 posterior probabilities using the direct estimation and logistic regression methods (Cornuet,
256 Pudlo, Veyssier, Dehne-Garcia, Gautier, Leblois, Marin & Estoup 2014a). We also
257 estimated the time for Out of Northern Kyushu using the Local dataset that included 698
258 SNPs from *DIYABC*. We inferred the following time-based simple evolutionary scenario:
259 N.Saga diverged from S.Saga at time t because this divergence event was consistent with the
260 Out of Northern Kyushu event (Katsumura et al. 2012). In this case, all populations are
261 assumed to have unchangeable effective sizes in each lineage.

262 The prior distribution of each parameter is presented in table 2, and the 1%
263 simulated datasets closest to the observed dataset were used to estimate the posterior

264 parameter distributions. We evaluated the accuracy of the demographic parameter
265 estimation by calculating accuracy indicators (*Bias and mean square error* option) in *DIYABC*
266 and accepted the parameters that were non-scaled and scaled by the mean effective population
267 size for the HZ-2 and Local datasets, respectively. Using the infinite-sites model in the
268 Wright-Fisher populations of a constant size, a rough value of N_e was estimated using the
269 relationship $\pi = 4N_e\mu$ (Tajima 1983), where π is genome-wide nucleotide diversity and μ is
270 the mutation rate per site per generation (Osada 2015). We calculated N_e using the above
271 formula with the following values: $\mu = 10^{-9}$/site/generation; generation = 1 year, which
272 resulted in a mean effective population size from all presented demes in northern Kyushu of
273 1,275,000.
274
275 **Reconstruction of the phylogenetic tree for the partial mitochondrial DNA sequence**
276 To examine the mitochondrial introgression in the tested samples, we generated and analyzed
277 the mitochondrial DNA (mtDNA) partial sequences from the GBS-read data. We mapped
278 the reads to the complete mtDNA sequence of "Clade C" (Matsuda et al. 1997; Takehana et al.
279 2003), which diverged from S.JPN and N.JPN, to eliminate the mapping bias that occurs
280 when genetically distant sequences are used as a reference. The genomic DNA of "Clade C"
281 was extracted in Katsumura *et al*. 2009. The complete mtDNA sequence was determined as
282 follows. Using PCR, five DNA fragments of mitochondrial DNA were amplified: fragment
283 1—4,556 bp; fragment 2—4,527 bp; fragment 3-1—4,589 bp; fragment 3-2—4,504 bp; and
284 fragment 4—4,546 bp (supplementary fig. 3). Supplementary table 1 describes the primers,
285 which were designed based on the inbred strain Hd-rR complete mtDNA sequence (accession
286 number: AP008938). Approximately 20 ng of the genomic DNA was used as a template for
287 the PCR assay in a 50 μl solution containing dNTP at 0.2 mM, 0.2 μM of each of primer, 0.75
288 U of EX Taq polymerase HS (TaKaRa Shuzo Co.), and the reaction buffer attached to the

289 polymerase. The reactions were conducted in a TaKaRa PCR Thermal Cycler Dice (TaKaRa
290 Shuzo Co.) using the following protocol: an initial denaturing step at 95°C for 2 min, 40
291 cycles of denaturation at 95°C for 30 sec, annealing at 60°C for 30 sec, extension at 72°C for
292 300 sec, and a final extension step at 72°C for 5 min. The PCR products were diluted 20
293 fold and used as templates in the sequencing reaction (following the commercial protocol)
294 with thirty-three primers (supplementary table 1, supplementary fig. 3) and then analyzed in
295 an ABI 3500xL Genetic Analyzer (Life Technologies). The complete mtDNA sequence of
296 "Clade C" was reconstructed using *SeqMan Pro 10.1.2.20* (DNASTAR) and deposited into
297 the international DNA database DDBJ/EMBL/GenBank (accession number: LC335803).
298 The "Clade C" complete mtDNA sequence was used to align the reads using *BWA*
299 *backtrack 0.7.12-r1039* (H Li & Durbin 2009) using the "-n 5" option. After the mapping
300 process, to remove the multi-mapped reads, we used *Samtools v1.2* (H Li et al. 2009) using
301 the "-Sq 20" option. Then, we used *Stacks* with the "-m 3" option for minimum stack depth
302 (Catchen et al. 2011; Catchen, Hohenlohe, et al. 2013) and then obtained the 322 bp
303 nucleotide sequences. The loci where the sequence was missing were filled with “N”. In
304 addition, the nucleotide position showing the multi-allelic state was replaced by “N”.
305 Phylogenetic trees were constructed using the neighbor-joining (NJ) method (Saitou & Nei
306 1987) with the program MEGA5 (Tamura et al. 2011). The evolutionary distances were
307 calculated using the Jukes-Cantor method (Jukes & Cantor 1969). The analysis also
308 involved 13 nucleotide sequences from the DNA database (See supplementary fig. 4). All
309 ambiguous positions were removed for each sequence pair. The reliability of the tree was
310 evaluated using 1000 bootstrap replicates (Felsenstein 1985).

311 Results

312 **Autosomal genetic diversity of five major groups of *Oryzias latipes***

313 To assess the genetic diversity based on autosomal SNPs within five geographical groups (fig.
314 1)—the northern Japanese (N.JPN), Tajima-Tango, southern Japanese (S.JPN), eastern
315 Korean (E.KOR) and western Korean/Chinese (W.KOR) groups—their population genetic
316 summary statistics were calculated using the "PopStat" dataset that included the alignment
317 regions across all groups and contained approximately 45 kb sequences and 2,453 SNPs
318 (table 1). All groups showed major allele frequencies and heterozygosities similar to the
319 geographic populations of wild-captured stickleback (Catchen, Bassham, et al. 2013),
320 indicating that the individuals in each group were genetically divergent from each other.
321 W.KOR showed the highest nucleotide diversity (0.0053 +/- 0.0003) in East Asia, and S.JPN
322 showed the highest nucleotide diversity (0.0036 +/- 0.0001) in the Japanese archipelago (table
323 1).

324

325 **Genetic relationship based on autosomal SNPs indicating medaka dispersal history**

326 To clarify the genetic relationship of medaka populations, we performed a PCA using the
327 "Global" dataset (fig. 2). First, we investigated whether the genetic relationship based on the
328 mitochondrial genome was reflected using 8,361 SNPs on 24 autosomes. The PCA showed
329 that the five groups were divided into two clusters, N.JPN/Tajima-Tango and S.JPN and
330 others from E/W.KOR, which is similar to the grouping based on the mtDNA analysis
331 (Takehana et al. 2003; Katsumura et al. 2009) (fig. 2a, supplementary fig. 5). E/W.KOR
332 individuals were plotted dispersedly compared with other clusters, and Yongcheon (E.KOR)
333 was close to N.JPN/Tajima-Tango.

334 Although Tajima-Tango was considered a hybridization group between N.JPN and
335 S.JPN (Sakaizumi 1984), all individuals from Tajima-Tango were plotted with those from

336 N.JPN as a cluster. Second, for a fine-scale mapping of Japanese medaka populations,
337 E/W.KOR individuals were removed from the analysis (fig. 2b). Based on 7,126 SNPs, the
338 PCA showed that Tajima-Tango overlapped with N.JPN again, whereas the Kyushu
339 populations in S.JPN were dispersed along the PC2 axis (supplementary fig. 6). These
340 analyses indicated that E.KOR and W.KOR were genetically distant from each other, and
341 Tajima-Tango was not considerably differentiated from N.JPN compared with S.JPN, which
342 was differentiated within the group.

343 To examine the phylogenetic relationship and the genomic ancestry–sharing among 83
344 populations, we constructed a maximum likelihood tree based on 4,638 short fragment
345 sequences (comparative nucleotide sequence length: 217,986 bp) using *IQ-TREE* and
346 performed a genetic clustering analysis using *ADMIXTURE*. The maximum likelihood tree
347 showed relatively high bootstrap values on each branch and almost the same topology as
348 previous trees based on mtDNA (Takehana et al. 2003; Katsumura et al. 2009) except for
349 Yongcheon (fig. 3 and supplementary fig. 7). Medakas from the Korean Peninsula were
350 genetically close to each other in both groups, whereas those from Shanghai, which have been
351 classified into W.KOR based on mtDNA studies, diverged from the other individuals from
352 W.KOR.

353 The medakas in the Japanese archipelago were divided into two major clusters. One
354 was the N.JPN/Tajima-Tango, and the other was S.JPN. The N.JPN/Tajima-Tango cluster
355 was further divided into two submajor clusters: N.JPN and Tajima-Tango. To compare the
356 genetic diversity among the mitochondrial groups, we calculated a mean pairwise branch
357 length within the geographic groups. N.JPN showed the lowest values (N.JPN: 0.0020,
358 Tajima-Tango: 0.0025, S.JPN: 0.0348, E.KOR: 0.0210, and W.KOR: 0.0120), indicating that
359 each local-wild population in the N.JPN was genetically homogeneous. S.JPN was also
360 further divided into two submajor clusters: Kyushu-only and Kyushu & others, which

361 included 14 Kyushu populations represented by an open and a closed red circle, respectively
362 (fig. 3).
363 While the Kyushu-only cluster diverged into the northern and the southern Kyushu,
364 the Kyushu & others cluster diverged to the Pacific side of eastern and the northern Honshu
365 (main-island Japan). The populations derived from northern Kyushu were positioned on the
366 most internal branches in both clusters, suggesting that northern Kyushu was the most likely
367 dispersal center for the Pacific-side medaka. This result strongly supports the "Out of
368 Northern Kyushu" hypothesis proposed based on nucleotide sequences of the mtD-loop
369 (Katsumura et al. 2012). In the northeastern Honshu cluster, the branch lengths were short
370 (mean pairwise branch length = 0.0032) and the bootstrap values were low, suggesting that
371 medakas were dispersed rapidly and were not yet highly differentiated in autosomes. With
372 this dataset, we characterized the S.JPN ancestry in the context of East Asian genetic diversity
373 by performing an *ADMIXTURE* analysis, which is a model-based unsupervised genetic
374 clustering method. With the optimal number of ancestral components (K = 4), S.JPN
375 medakas were assigned to two distinct ancestries (supplementary fig. 2 and fig. 3). In
376 suboptimal runs with more ancestral components (K = 5, 6), only S.JPN was assigned to the
377 other ancestral components found in the northern Kyushu populations. This analysis showed
378 that genetic diversities in S.JPN originated from the northern Kyushu populations. Thus, the
379 phylogenetic analyses based on genome-wide SNP data strongly supported that S.JPN spread
380 from northern Kyushu to Honshu.
381
382 **Boundary population genomes similar to that of the northern Japanese group**
383 To examine allele-sharing between boundary populations and surrounding populations using
384 the "HZ-1" dataset (see Materials & Methods), we classified the fixed alleles between the
385 groups into three states: shared by N.JPN and Tajima-Tango, shared by S.JPN and

386 Tajima-Tango, and shared by N.JPN and S.JPN. Then, we summarized each state's total
387 numbers as shown in table in fig. 4. We found 1,380 out of 4,661 SNPs that were common
388 alleles between two of the three groups (fig. 4). Regarding those alleles, the majority were
389 the common alleles between Tajima-Tango and N.JPN at 81.4%, which was a much higher
390 frequency than that between Tajima-Tango and S.JPN (11.2%). The remainder were the
391 common alleles between N.JPN and S.JPN (7.5%), i.e., specific alleles in Tajima-Tango.
392 The proportions are nearly consistent with the previous study based on 96 genomic regions (4
393 loci for each chromosome) (Takehana et al. 2016). Although the regions with neighboring
394 common alleles between Tajima-Tango and S.JPN were observed for certain chromosomes,
395 alleles on the Tajima-Tango genome were mostly shared with N.JPN, indicating that
396 Tajima-Tango is a subgroup of N.JPN with a short divergence time.
397 To investigate the mitochondria introgression in boundary populations, we
398 reconstructed the phylogenetic tree based on partial mtDNA sequences generated by mapping
399 short reads from the GBS of the complete mitochondrial genome (supplementary fig. 4).
400 The mtDNA sequences from Toyooka and Kinosaki in Tajima-Tango were clustered and
401 closely related to those from Kaga, which was a root population in N.JPN. Considering that
402 Tajima-Tango is a subgroup of N.JPN inferred from the tree based on autosomal sequences,
403 these phylogenetic positions on the mtDNA tree would have reflected an evolutionary history
404 in which N.JPN diverged from the Tajima-Tango group. We confirmed that Amino,
405 Kumihama and Hamasaka in Tajima-Tango had the mitochondrial genomes of
406 phylogenetically distant populations classified into S.JPN (supplementary fig. 4).
407 Additionally, the mitochondrial genome of Ayabe in S.JPN was classified into N.JPN but
408 diverged from Tajima-Tango. These results suggest that mitochondrial genome
409 introgressions occurred reciprocally, meaning that they occurred not only from S.JPN to
410 Tajima-Tango but also from N.JPN to S.JPN.

411

412 **Inferring medaka demographic parameters**

413 To infer the demographic parameters, effective population size (N_e), time (T), and proportion
414 of the admix (r), we analyzed the "HZ-2" dataset based on the coalescent theory using an
415 Approximate Bayesian Computation (ABC) framework. We performed the model selection
416 to identify the best explanation scenario for the observed data from the four scenarios (fig.
417 5a).

418 *DIYABC* has two different approaches (directional and logistic) for model selection.
419 Based on each criterion, scenario III (Honshu diverged from Kyushu and then N.JPN diverged
420 from the Tajima-Tango group without admixture with Honshu) and scenario IV (Honshu
421 diverged from Kyushu and then Tajima-Tango occurred by an admixture with Honshu at r
422 rate) were supported by the directional (posterior probability: 0.4720, 95% CI: 0.0344–
423 0.9096) and logistic (posterior probability: 0.4477, 95% CI: 0.4313–0.4641) approaches,
424 respectively (supplementary fig. 8). Unfortunately, our data could not distinguish one of
425 these two scenarios (see the Discussion section).

426 The posterior parameter estimates of scenarios III and IV are shown in table 2 and
427 figs. 5c and 5d, which were not scaled because various measures of accuracy (RRMISE,
428 RMeanAD, and RRMSE in the *DIYABC* output) indicated that non-scaled parameters fit the
429 observed data better than scaled parameters Additionally, although we inferred the time to
430 the most common ancestor at three divergence events (fig. 5), these times might be
431 overestimated because the dataset for this estimation consisted of non-deme samples, i.e.,
432 polymorphisms within populations were underestimated and the random-mating assumption
433 could not be satisfied. Therefore, we estimated the timing of the Out of Northern Kyushu
434 event using the "Local" dataset composed of the two deme samples (S.Saga and N.Saga),
435 which were split at first on the S.JPN lineage (fig. 3). Based on a simple hypothesis,

436 constant population size and no migration, we gained the estimated time (510,000 years ago,
437 95% CI: 337,875–679,575) for the ancestral divergence of the two deme samples (table 2 and
438 fig. 5e), which was calculated from a scaled parameter by the mean effective population size
439 because various accuracy measures indicated that scaled parameters fit the observed data
440 better than non-scaled parameters. Thus, our population genetic estimate using ABC
441 suggests that medakas diverged and dispersed throughout Pacific-side Japan at approximately
442 510 kyr ago.

443 Discussion

444 **Genetic diversity and structure based on genome-wide single nucleotide polymorphisms**

445 The previous study showed that N.JPN and S.JPN could be divided into three and twelve
446 subgroups, respectively, according to the mitochondrial DNA analysis (Takehana et al. 2003).
447 In this study, based on autosomal SNPs, the genetic clustering analysis showed that "K = 4"
448 was the most supportive because it presented the lowest fivefold cross-validation error,
449 suggesting that N.JPN and S.JPN were divided into three ancestral clusters. When the K
450 values increased, only S.JPN divided into more subgroups, which suggests that the S.JPN
451 group was composed of more divergent groups than the other groups. S.JPN can be divided
452 into finer subgroups based on the mitochondrial genome, which is likely because the effective
453 population size of mtDNA is one quarter that of a nuclear gene. Although this result does
454 not indicate that the autosomal genome has been less differentiated, it does suggest that the
455 autosomal genome in each population shares a common ancestor.

456 The medaka mtDNA tree shows that the intermingled branch pattern is not
457 associated with geographic distance. This pattern is similar to that based on allozyme data
458 shown in Sakaizumi *et al.* (1986), which has been interpreted as artificial migrations
459 accompanied by recent human activities (Takehana et al. 2003). If the branching pattern was
460 formed by human activities, various ancestral components should appear independent of
461 geographic relatedness in the result of the *ADMIXTURE* analysis based on autosomal SNPs.
462 However, no signal of recent migration was found in the *ADMIXTURE* result. Rather, the
463 autosomal SNP data support another hypothetical scenario in which the mtDNA tree topology
464 reflected ancestral polymorphisms only and their local fixation was caused by a small
465 effective population size (Katsumura et al. 2009; 2012). This finding indicates that the wild
466 lab stocks still maintain their migration memory on their genomes.

467 Among four major mitochondrial groups, W.KOR showed the highest nucleotide
468 diversity and E.KOR showed the second highest (table 1). The former group included the
469 Chinese medaka in Shanghai, which could elevate the value, while the latter group did not
470 include any geographically distant populations. The clustering analysis showed that
471 Shanghai from W.KOR and Yongcheon from E.KOR had the ancestry component from N.JPN,
472 which suggests two possibilities: one is that the N.JPN ancestor derives from the E/W.KOR
473 ancestor, and the other is that contamination occurred through the maintained wild lab stocks.
474 To investigate these possibilities and the origin of *O. latipes*, a population-based genome wide
475 analysis must be conducted to increase the population numbers in E/W.KOR and include the
476 sister species *O. curvinotus* and *O. luzonensis*.
477
478 **Reconstruction of the medaka population history in the Japanese archipelago**
479 Our genome-wide analysis shows that medakas in N.JPN and S.JPN are deeply divergent and
480 dispersed over the Japanese archipelago from different locations at different times. In
481 particular, the history of N.JPN after divergence from S.JPN cannot be consistent with
482 previous inferences from the allozyme, mtDNA and limited autosomal SNPs analyses, which
483 have suggested Tajima-Tango originated in and diverged from N.JPN. Our data and
484 previous studies show that (i) the nucleotide diversity in Tajima-Tango is higher than that in
485 N.JPN (table 1), (ii) S.JPN is genetically more closely related to Tajima-Tango than to N.JPN
486 based on the allele-sharing rates (fig. 4), and (iii) the Tajima-Tango branch is the root in the
487 N.JPN/Tajima-Tango clade of the phylogenetic tree based on partial mtDNA sequences (fig.
488 S4) and the whole mitochondrial genome (Hirayama et al. 2010). These data suggest that
489 Tajima-Tango forms an outgroup to all present-day N.JPN and it spread along the
490 Sea-of-Japan side. Furthermore, the ABC framework's estimation supports our scenario,
491 although the analysis does not statistically distinguish between the two scenarios (admixture

492 or divergent). Even if the admixture occurred, the inferred ratio of the admixture is too low
493 (fig. 5d).
494 In this study, we support the hypothetical scenario "Out of northern Kyushu model,"
495 which was first reported by Katsumura et al. (2012), and discover the dispersal route of
496 N.JPN/Tajima-Tango. The genetic clustering analyses and phylogenetic tree based on the
497 GBS data elucidate medaka history better than previous mitochondrial DNA analyses. In
498 particular, the *ADMIXTURE* analysis shows that the two ancestry components (yellow and
499 blue in fig. 3) were observed in northern Kyushu, suggesting that S.JPN dispersed in three
500 different directions after dispersing Out of Northern Kyushu. Furthermore, the different
501 branching patterns of autosomal and mtDNA trees suggest that the mtDNA introgression
502 occurred not only from S.JPN to Tajima-Tango but also from N.JPN to S.JPN. The
503 Tajima-Tango region is surrounded by mountains (supplementary fig. 1); however, it contains
504 the lowest watershed (sea level 95.45 m) in the Japanese archipelago. The medaka has
505 possibly moved in both directions across the watershed. From the above, the most plausible
506 scenario is as follows (fig. 6). S.JPN and Tajima-Tango diverged first and independently
507 reached their current habitats in the Japanese archipelago. After S.JPN dispersed from
508 northern Kyushu approximately 510 kyr ago, N.JPN diverged from Tajima-Tango. Then, their
509 descendant populations spread rapidly to the northern Honshu on the Sea-of-Japan side.
510 Meanwhile, S.JPN may have dispersed in three different directions and then spread rapidly
511 northeastward from the western part of Fossa Magna. In the process of dispersing across the
512 main island of Japan, certain S.JPN populations infiltrated the Tajima-Tango region from the
513 west and the south, and the resulting mtDNA introgression occurred independently.
514
515 **Population structure and history information for the medaka GWAS**
516 The population structure inferred from autosomes helps us select the optimal sampling

517 strategy for performing a GWAS. The proportion of false positives and false negatives in
518 the GWAS increases when the population structure overlaps with the phenotypic pattern
519 (Brachi et al. 2011). To avoid this result, the population structure and history of organisms
520 used for a GWAS should be examined based on many genomic regions with recombination.
521 Although our medaka SNP data are few compared with the amount of available human SNP
522 data, these data provide a baseline genetic diversity estimate for the detection of fine-scale
523 population structures and reconstruction of the population history compared with the
524 mitochondrial genome. With regard to the GWAS with the medaka population (medaka
525 GWAS), this study can be useful for identifying a relationship between the phenotypic pattern
526 and population structure and planning a sampling strategy to avoid false positives/negatives.
527 The explanatory rate for observed SNP traits is low in many cases in the human GWAS,
528 whereas with model organisms, particularly plants, GWAS results have often explained a
529 much greater proportion of the phenotypic variation (Yan Li et al. 2010). Therefore, the
530 assumption holds true for such organisms that common genetic variation explains common
531 phenotypic variation. Therefore, many adaptive traits and related genes should be identified
532 via the medaka GWAS while considering the population structure. In addition, cutting-edge
533 techniques for genome editing, such as CRISPR/cas9 (Kirchmaier et al. 2015), and resource
534 management systems, such as NBRP medaka (URL: https://shigen.nig.ac.jp/medaka/) have
535 been established. Whole genome analyses of inbred lines and some wild-captured medakas
536 have already been performed. If autosomal information can be accumulated for wild lab
537 stocks, medakas may become the only model animal used for the systematic performance of
538 GWASs, which is similar to the scenario of Arabidopsis for plant GWASs. Thus, medakas
539 have the potential for use as a highly sensitive tool for unraveling the molecular mechanism
540 of phenotypic variations and the adaptive significance of genetic polymorphisms in a
541 population, providing a useful key for understanding the evolution of phenotypic

542 diversification in vertebrates.

543

544 Reference

545 Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry in
546 unrelated individuals. Genome Research. 19:1655–1664. doi: 10.1101/gr.094052.109.

547 Andrews KR, Good JM, Miller MR, Luikart G, Hohenlohe PA. 2016. Harnessing the power
548 of RADseq for ecological and evolutionary genomics. Nature Review Genetics. 17:81–92.
549 doi: 10.1038/nrg.2015.28.

550 Ansai S, Kinoshita M. 2014. Targeted mutagenesis using CRISPR/Cas system in medaka.
551 Biology Open. 3:362–371. doi: 10.1242/bio.20148177.

552 Beaumont MA, Zhang W, Balding DJ. 2002. Approximate Bayesian computation in
553 population genetics. Genetics. 162:2025–2035.

554 Brachi B, Morris GP, Borevitz JO. 2011. Genome-wide association studies in plants: the
555 missing heritability is in the field. Genome Biol. 12:232. doi: 10.1186/gb-2011-12-10-232.

556 Catchen J, Bassham S, et al. 2013. The population structure and recent colonization history of
557 Oregon threespine stickleback determined using restriction-site associated DNA-sequencing.
558 Mol Ecol. 22:2864–2883. doi: 10.1111/mec.12330.

559 Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA. 2013. Stacks: an analysis tool
560 set for population genomics. Mol Ecol. 22:3124–3140. doi: 10.1111/mec.12354.

561 Catchen JM, Amores A, Hohenlohe P, Cresko W, Postlethwait JH. 2011. Stacks: building and
562 genotyping Loci de novo from short-read sequences. De Koning, D-J, editor. G3 (Bethesda).
563 1:171–182. doi: 10.1534/g3.111.000240.

564 Chang CC et al. 2015. Second-generation PLINK: rising to the challenge of larger and richer
565 datasets. GigaSci. 4:7–16. doi: 10.1186/s13742-015-0047-8.

566 Chernomor O, Haeseler von A, Minh BQ. 2016. Terrace Aware Data Structure for
567 Phylogenomic Inference from Supermatrices. Systematic Biology. 65:997–1008. doi:
568 10.1093/sysbio/syw037.

569 Consortium T1G et al. 2016. 1,135 Genomes Reveal the Global Pattern of Polymorphism in
570 Arabidopsis thaliana. Cell. 166:481–491. doi: 10.1016/j.cell.2016.05.063.

571 Cornuet J-M, Pudlo P, Veyssier J, Dehne-Garcia A, Gautier M, Leblois R, Marin J-M, Estoup
572 A. 2014a. DIYABC v2.0: a software to make approximate Bayesian computation inferences
573 about population history using single nucleotide polymorphism, DNA sequence and
574 microsatellite data. Bioinformatics. 30:1187–1189. doi: 10.1093/bioinformatics/btt763.

575 Cornuet J-M, Pudlo P, Veyssier J, Dehne-Garcia A, Gautier M, Leblois R, Marin J-M, Estoup
576 A. 2014b. DIYABC v2.0: a software to make approximate Bayesian computation inferences
577 about population history using single nucleotide polymorphism, DNA sequence and
578 microsatellite data. Bioinformatics. 30:1187–1189. doi: 10.1093/bioinformatics/btt763.

579 Elshire RJ et al. 2011. A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for
580 High Diversity Species Orban, L, editor. PLoS ONE. 6:e19379–10. doi:
581 10.1371/journal.pone.0019379.

582 Felsenstein J. 1985. Confidence limits on phylogenies: an approach using the bootstrap.
583 Evolution. 783–791.

584 Gonzales NM, Palmer AA. 2014. Fine-mapping QTLs in advanced intercross lines and other
585 outbred populations. Mamm Genome. 25:271–292. doi: 10.1007/s00335-014-9523-1.

586 Guindon S et al. 2010. New Algorithms and Methods to Estimate Maximum-Likelihood
587 Phylogenies: Assessing the Performance of PhyML 3.0. Systematic Biology. 59:307–321.
588 doi: 10.1093/sysbio/syq010.

589 Hirayama M et al. 2010. Intraspecific variation in the mitochondrial genome among local
590 populations of Medaka Oryzias latipes. Gene. 457:13–24. doi: 10.1016/j.gene.2010.02.012.

591 Igarashi K et al. 2017. An Approach to Elucidate NBS1 Function in DNA Repair Using
592 Frequent Nonsynonymous Polymorphism in Wild Medaka (Oryzias latipes) Populations
593 Fugmann, SD, editor. PLoS ONE. 12:e0170006–19. doi: 10.1371/journal.pone.0170006.

594 Jeong C, Nakagome S, Di Rienzo A. 2016. Deep history of East Asian populations revealed
595 through genetic analysis of the Ainu. Genetics. doi: 10.1534/genetics.115.178673/-/DC1.

596 Jukes TH, Cantor CR. 1969. *Evolution of protein molecules. In 'Mammalian Protein*
597 *Metabolism'.(Ed. HN Munro.) pp. 21–132*.

598 Katsumura T et al. 2012. A population genetic study on the relationship between medaka fish
599 and the spread of wet-rice cultivation across the Japanese archipelago. Anthropol. Sci.

600 Katsumura T et al. 2009. Genetic differentiation among local populations of medaka fish
601 (Oryzias latipes) evaluated through grid- and deme-based sampling. Gene. 443:170–177. doi:
602 10.1016/j.gene.2009.04.017.

603 Katsumura T et al. 2014. Natural allelic variations of xenobiotic-metabolizing enzymes affect
604 sexual dimorphism in Oryzias latipes. Proceedings of the Royal Society B: Biological
605 Sciences. 281. doi: 10.1098/rspb.2014.2259.

606 Kirby A et al. 2010. Fine Mapping in 94 Inbred Mouse Strains Using a High-Density
607 Haplotype Resource. Genetics. 185:1081–1095. doi: 10.1534/genetics.110.115014.

608 Kirchmaier S, Naruse K, Wittbrodt J, Loosli F. 2015. The Genomic and Genetic Toolbox of
609 the Teleost Medaka (Oryzias latipes). Genetics. 199:905–918. doi:
610 10.1534/genetics.114.173849.

611 Li H et al. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics.
612 25:2078–2079. doi: 10.1093/bioinformatics/btp352.

613 Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler
614 transform. Bioinformatics. 25:1754–1760. doi: 10.1093/bioinformatics/btp324.

615 Li Y, Huang Y, Bergelson J, Nordborg M, Borevitz JO. 2010. Association mapping of local
616 climate-sensitive quantitative trait loci in Arabidopsis thaliana. Proceedings of the National
617 Academy of Sciences. 107:21199–21204. doi: 10.1073/pnas.1007431107.

618 Mackay TFC et al. 2012. The Drosophila melanogaster Genetic Reference Panel. Nature.
619 482:173–178. doi: 10.1038/nature10811.

620 Matsuda M, Yonekawa H, Hamaguchi S, Sakaizumi M. 1997. Geographic variation and
621 diversity in the mitochondrial DNA of the medaka, Oryzias latipes, as determined by
622 restriction endonuclease analysis. Zool. Sci. 14:517–526.

623 Minh BQ, Nguyen MAT, Haeseler von A. 2013. Ultrafast Approximation for Phylogenetic
624 Bootstrap. Molecular Biology and Evolution. 30:1188–1195. doi: 10.1093/molbev/mst024.

625 Narum SR, Buerkle CA, Davey JW, Miller MR, Hohenlohe PA. 2013.
626 Genotyping-by-sequencing in ecological and conservation genomics. Mol Ecol. 22:2841–
627 2847. doi: 10.1111/mec.12350.

Nguyen LT, Schmidt HA, Haeseler von A, Minh BQ. 2014. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. Molecular Biology and Evolution. 32:268–274. doi: 10.1093/molbev/msu300.

Oota H, Mitani H. 2010. *Human Population Genetics Meets Medaka*.

Osada N. 2015. Genetic diversity in humans and non-human primates and its evolutionary consequences. Genes Genet. Syst. 90:133–145. doi: 10.1266/ggs.90.133.

Parker CC et al. 2016. Genome-wide association study of behavioral, physiological and gene expression traits in outbred CFW mice. Nat Genet. 48:919–926. doi: 10.1038/ng.3609.

Poland JA, Brown PJ, Sorrells ME, Jannink J-L. 2012. Development of High-Density Genetic Maps for Barley and Wheat Using a Novel Two-Enzyme Genotyping-by-Sequencing Approach Yin, T, editor. PLoS ONE. 7:e32253–8. doi: 10.1371/journal.pone.0032253.

Saitou N, Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular Biology and Evolution. 4:406–425.

Sakaizumi M. 1986. Genetic divergence in wild populations of Medaka, Oryzias latipes (Pisces: Oryziatidae) from Japan and China. Genetica.

Sakaizumi M. 1984. Rigid Isolation between the Northern Population and the Southern Population of the Medaka, Oryzias latipes(Genetics). Zool. Sci. 1:795–800.

Sakaizumi M, EGAMI N, MORIWAKI K. 1980. Allozymic Variation in Wild Populations of the Fish, Oryzias latipes. Proc. Jpn. Acad., Ser. B. 56:448–451. doi: 10.2183/pjab.56.448.

Sakaizumi M, Moriwaki K, Egami N. 1983. Allozymic variation and regional differentiation in wild populations of the fish Oryzias latipes. Copeia. 1983:311. doi: 10.2307/1444373.

649 Shima A, Shimada A, Sakaizumi M, Egami N. 1985. First listing of wild stocks of the
650 medaka Oryzias latipes currently kept by the Zoological Institute, Faculty of Science,
651 University of Tokyo. Journal of the Faculty of Science, Imperial University of Tokyo Sect IV,
652 Zoology. 16:27–35.

653 Spivakov M et al. 2014. Genomic and phenotypic characterization of a wild medaka
654 population: towards the establishment of an isogenic population genetic resource in fish. G3
655 (Bethesda). 4:433–445. doi: 10.1534/g3.113.008722.

656 Tajima F. 1983. Evolutionary relationship of DNA sequences in finite populations. Genetics.
657 105:437–460.

658 Takehana Y et al. 2016. Origin of Boundary Populations in Medaka (Oryzias latipes Species
659 Complex). Zool. Sci. 33:125–131. doi: 10.2108/zs150144.

660 Takehana Y, Nagai N, Matsuda M, Tsuchiya K, Sakaizumi M. 2003. Geographic variation
661 and diversity of the cytochrome b gene in Japanese wild populations of medaka, Oryzias
662 latipes. Zool. Sci. 20:1279–1291. doi: 10.2108/zsj.20.1279.

663 Tamura K et al. 2011. MEGA5: Molecular Evolutionary Genetics Analysis Using Maximum
664 Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. Molecular Biology
665 and Evolution. 28:2731–2739. doi: 10.1093/molbev/msr121.

666 Watanabe-Asaka T et al. 2014. Regular heartbeat rhythm at the heartbeat initiation stage is
667 essential for normal cardiogenesis at low temperature. BMC Dev. Biol. 14. doi:
668 10.1186/1471-213X-14-12.

669 Welter D et al. 2013. The NHGRI GWAS Catalog, a curated resource of SNP-trait
670 associations. Nucleic Acids Research. 42:D1001–D1006. doi: 10.1093/nar/gkt1229.

671 Wickham H. 2011. ggplot2. Wiley Interdisciplinary Reviews: Computational Statistics.
672 3:180–185. doi: 10.1002/wics.147.

673 Zheng X et al. 2012. A high-performance computing toolset for relatedness and principal
674 component analysis of SNP data. Bioinformatics. 28:3326–3328. doi:
675 10.1093/bioinformatics/bts606.

676

Acknowledgments

We thank Mrs. Shizuko Chiba, Mrs. Sumiko Tomizuka, Dr. Atsuko Shimada and Prof. Emeritus Akihiro Shima (the University of Tokyo) for maintaining the medaka stocks from wild populations. This study was supported in part by grants-in-aid for Young Scientists (B), nos. 16K21352 to T.K. T.K. was also supported by a grant-in-aid from JSPS Research Fellow (16J07227).

Table 1 Summary genetic statistics for five populations using "PopStat" dataset. These statistics include the mean number of individuals genotyped at each locus (*1), the number of variable sites in five populations (Variants), the number of polymorphic sites in each population (Polymorphic) and the number of variable sites unique to each population (Private). The number in parenthesis is Standard Error.

Group	Number of Individual[*1]	Total length of GBS loci (bp)	Variant	Polymorphic	Private	Major Allele Frequency	Observed Heterozygosity	Nucleotide Diversity
NJPN	11.5 (+/-0.03)	45968	2453	256	136	0.983 (+/-0.0014)	0.023 (+/-0.0022)	0.0014 (+/-0.0001)
Tajima -Tango	4.4 (+/-0.01)	44011	2453	199	71	0.980 (+/-0.0016)	0.024 (+/-0.0022)	0.0017 (+/-0.0001)
SJPN	41.6 (+/-0.08)	45968	2453	1448	1231	0.958 (+/-0.0017)	0.029 (+/-0.0016)	0.0036 (+/-0.0001)
WKOR	2.4 (+/-0.01)	44990	2453	467	342	0.940 (+/-0.0027)	0.054 (+/-0.0032)	0.0053 (+/-0.0003)
EKOR	2.4 (+/-0.01)	44011	2453	346	203	0.953 (+/-0.0025)	0.041 (+/-0.0030)	0.0042 (+/-0.0002)

Table 2 Demographic parameters estimated by DIYABC under three scenarios. Performance of each estimation was evaluated by RRMISE (the square Root of the Relative Mean Integrated Square Error), RMeanAD (the Relative Mean Absolute Deviation) and RRMSE (the square Root of teh Relative Mean Square Error), which are output from DIYABC option "Compute bias and mean square error." The numbers with or without parentheses in the culunms "Performances for estimating posterior distributions of parameters" are those of the statistics computed from the prior or posterior distribution of parameters, respectively.

		Prior distribution					Posterior parameter estimates			Performances for estimating posterior distributions of parameters		
Scenario	Parameters	Type	min.	max.	mean	S.D.	Mean	95% credible interval		RRMISE	RMeanAD	RRMSE
II	$N_{\text{N.JPN}}$	Normal	10,000	3,000,000	450,000	500,000	825,000	492,000	1,240,000	0.415 (-0.586)	0.302 (-0.485)	0.302 (-0.288)
	$N_{\text{Tajima-Tango}}$	Normal	10,000	3,000,000	600,000	500,000	613,000	369,000	901,000	0.35 (-0.808)	0.256 (-0.61)	0.251 (-0.358)
	N_{Kyushu}	Normal	10,000	3,000,000	1,225,000	500,000	1,210,000	711,000	1,830,000	0.364 (-0.524)	0.264 (-0.396)	0.267 (-0.282)
	N_{honshu}	Normal	10,000	3,000,000	1,050,000	500,000	1,490,000	974,000	2,020,000	0.291 (-0.442)	0.219 (-0.368)	0.213 (-0.289)
	N_{Ancestor}	Uniform	10,000	30,000,000	-	-	432,000	59,700	922,000	5.899 (-133.368)	1.846 (-59.034)	5.119 (-115.475)
	$T_{\text{N.JPN–Hybrid}}$	Uniform	10	4,000,000	-	-	566,000	311,000	908,000	0.409 (-2.417)	0.300 (-1.725)	0.266 (-1.589)
	$T_{\text{kyushu–Honshu}}$	Uniform	10	4,000,000	-	-	1,050,000	442,000	2,290,000	0.575 (-2.398)	0.374 (-1.875)	0.284 (-2.114)
	$T_{\text{N.JPN–S.JPN}}$	Normal	1,000,000	30,000,000	4,000,000	5,000,000	4,560,000	2,820,000	6,810,000	0.329 (-1.079)	0.245 (-0.776)	0.229 (-0.743)
III	$N_{\text{N.JPN}}$	Normal	10,000	3,000,000	450,000	500,000	812,000	390,000	1,330,000	0.447 (-0.635)	0.332 (-0.506)	0.303 (-0.326)
	$N_{\text{Tajima-Tango}}$	Normal	10,000	3,000,000	600,000	500,000	480,000	266,000	739,000	0.375 (-1.208)	0.284 (-0.88)	0.261 (-0.736)
	N_{Kyushu}	Normal	10,000	3,000,000	1,225,000	500,000	1,290,000	758,000	1,910,000	0.356 (-0.483)	0.265 (-0.373)	0.265 (-0.252)
	N_{honshu}	Normal	10,000	3,000,000	1,050,000	500,000	1,450,000	913,000	2,020,000	0.316 (-0.446)	0.235 (-0.369)	0.235 (-0.279)
	N_{Ancestor}	Uniform	10,000	30,000,000	-	-	390,000	57,200	849,000	6.947 (-131.19)	2.285 (-62.259)	5.997 (-113.55)
	T_{ADMIX}	Uniform	10	4,000,000	-	-	512,000	242,000	842,000	0.397 (-4.374)	0.308 (-2.163)	0.249 (-3.309)
	$T_{\text{kyushu–Honshu}}$	Uniform	10	4,000,000	-	-	1,100,000	446,000	2,420,000	0.576 (-2.256)	0.381 (-1.76)	0.296 (-1.98)
	$T_{\text{N.JPN–S.JPN}}$	Normal	1,000,000	30,000,000	4,000,000	5,000,000	5,160,000	3,080,000	7,780,000	0.345 (-0.916)	0.262 (-0.65)	0.239 (-0.588)
	r	Uniform	0.001	0.999	-	-	0.944	0.897	0.987	0.027 (-0.559)	0.02 (-0.472)	0.014 (-0.469)
Out of NK	$N_{\text{S.Saga}}$	Normal	10,000	3,000,000	1,300,000	500,000	1,326,000	960,075	1,670,250	0.178 (-0.31)	0.138 (-0.241)	0.120 (-0.095)
	$N_{\text{N.Saga}}$	Normal	10,000	3,000,000	1,250,000	500,000	1,224,000	892,500	1,593,750	0.181 (-0.322)	0.142 (-0.251)	0.119 (-0.096)
	$T_{\text{S.Saga–N.Saga}}$	Uniform	1,000	4,000,000	-	-	510,000	337,875	679,575	0.224 (-5.918)	0.175 (-4.488)	0.143 (-4.429)

683 Figure legends

684 **FIG. 1: Map of original locations of the wild lab stocks and wild-captured medakas.** In
685 the upper central map, the enlarged red frame shows the boundary region between S.JPN and
686 N.JPN. Each color represents the geographical groups shown in the left upper inset box.
687 The numbers on the map are consistent with the population IDs with the names on the right
688 bottom inset box.

689

690 **FIG. 2: Results of the principle component analysis (PCA) using SNPs in East Asia (a)**
691 **and the Japanese archipelago (b).** Each plot shows PC1 versus PC2. The population
692 names of each point are described in fig. S6.

693

694 **FIG. 3: Phylogenetic tree using the maximum likelihood method and an ancestry**
695 **barplot with an ADMIXTURE analysis.** Red closed and open circles represent the
696 northern and southern Kyushu populations, respectively. “Taj-Tan” in the tree is the
697 abbreviation for Tajima-Tango.

698

699 **FIG. 4: Shared allele distribution among boundary populations.** The gray rectangle
700 represents the total length of each medaka chromosome. The table in the figure shows the
701 number of alleles shared between the groups seen on each chromosome. Mark represents
702 the state whose group shares that allele. *1 and *2 are the total number from chromosomes
703 1 to 12 and from chromosomes 13 to 24, respectively.

704

705 **FIG. 5: Scenarios for the estimation of demographic parameters.** Possible scenarios for
706 the population history of the Tajima-Tango group (a), and the "Out of northern Kyushu (NK)"
707 event (b) are shown. The TMRCAs estimated using ABC are in scenario3 (c), scenario4 (d)

708 and Out of NK (e).

709

710 **FIG. 6: Map representing the ancestry proportions from the ADMIXTURE analysis at**

711 **K = 6.** Solid and dashed lines represent the spreading patterns of S.JPN and N.JPN inferred

712 by GBS data, respectively.

Figure 1

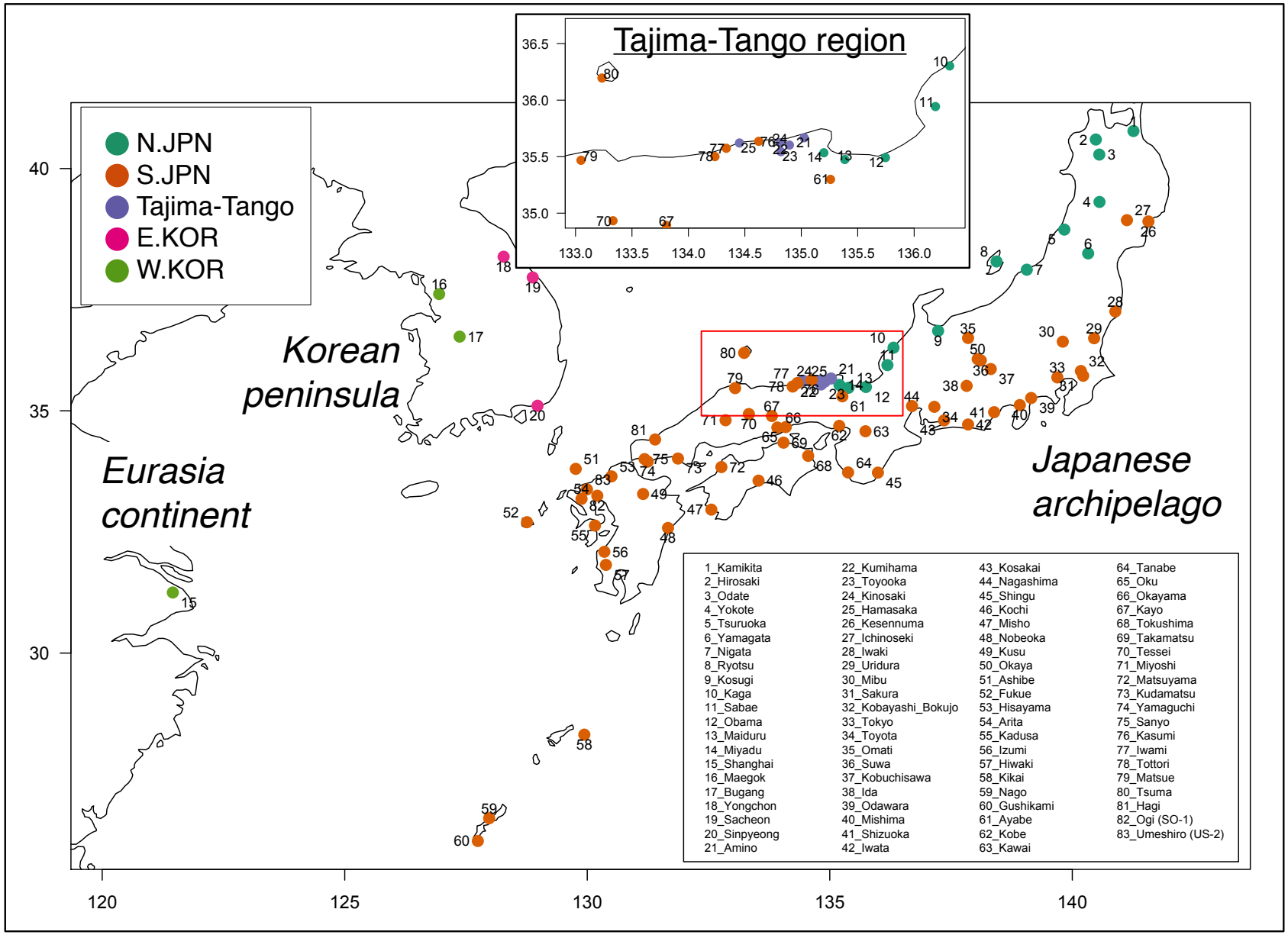

Tajima-Tango region
N.JPN
S.JPN
Tajima-Tango
E.KOR
W.KOR
Korean peninsula
Eurasia continent
Japanese archipelago
1_Kamikita
2_Hirosaki
3_Odate
4_Yokote
5_Tsuruoka
6_Yamagata
7_Nigata
8_Ryotsu
9_Kosugi
10_Kaga
11_Sabae
12_Obama
13_Maiduru
14_Miyadu
15_Shanghai
16_Maegok
17_Bugang
18_Yongchon
19_Sacheon
20_Sinpyeong
21_Amino
22_Kumihama
23_Toyooka
24_Kinosaki
25_Hamasaka
26_Kesennuma
27_Ichinoseki
28_Iwaki
29_Uridura
30_Mibu
31_Sakura
32_Kobayashi_Bokujo
33_Tokyo
34_Toyota
35_Omati
36_Suwa
37_Kobuchisawa
38_Ida
39_Odawara
40_Mishima
41_Shizuoka
42_Iwata
43_Kosakai
44_Nagashima
45_Shingu
46_Kochi
47_Misho
48_Nobeoka
49_Kusu
50_Okaya
51_Ashibe
52_Fukue
53_Hisayama
54_Arita
55_Kadusa
56_Izumi
57_Hiwaki
58_Kikai
59_Nago
60_Gushikami
61_Ayabe
62_Kobe
63_Kawai
64_Tanabe
65_Oku
66_Okayama
67_Kayo
68_Tokushima
69_Takamatsu
70_Tessei
71_Miyoshi
72_Matsuyama
73_Kudamatsu
74_Yamaguchi
75_Sanyo
76_Kasumi
77_Iwami
78_Tottori
79_Matsue
80_Tsuma
81_Hagi
82_Ogi (SO-1)
83_Umeshiro (US-2)

Figure 2

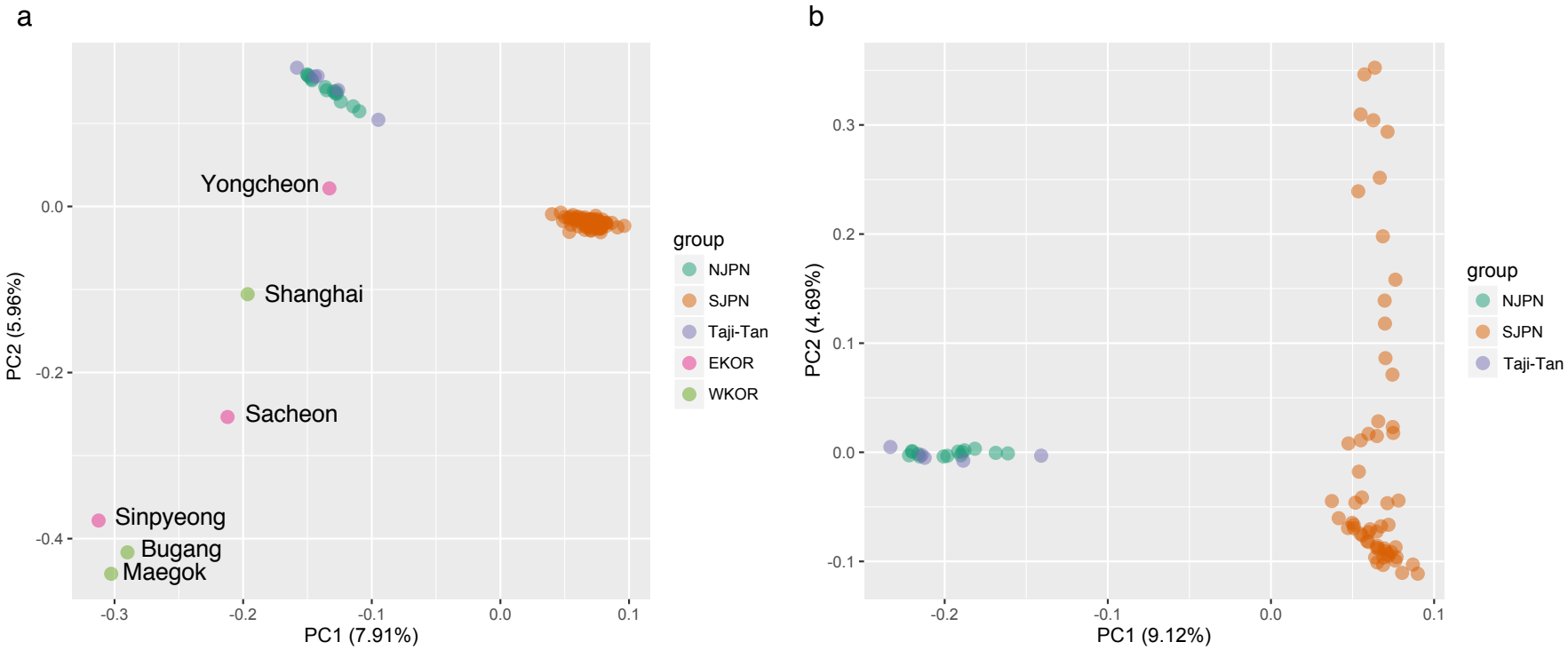

a
Yongcheon
Shanghai
Sacheon
Sinpyeong
Bugang
Maegok
PC2 (5.96%)
PC1 (7.91%)
0.0
-0.2
-0.4
-0.3
-0.2
-0.1
0.0
0.1
group
NJPN
SJPN
Taji-Tan
EKOR
WKOR
b
PC2 (4.69%)
PC1 (9.12%)
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.1
0.0
0.1
group
NJPN
SJPN
Taji-Tan

Figure 3

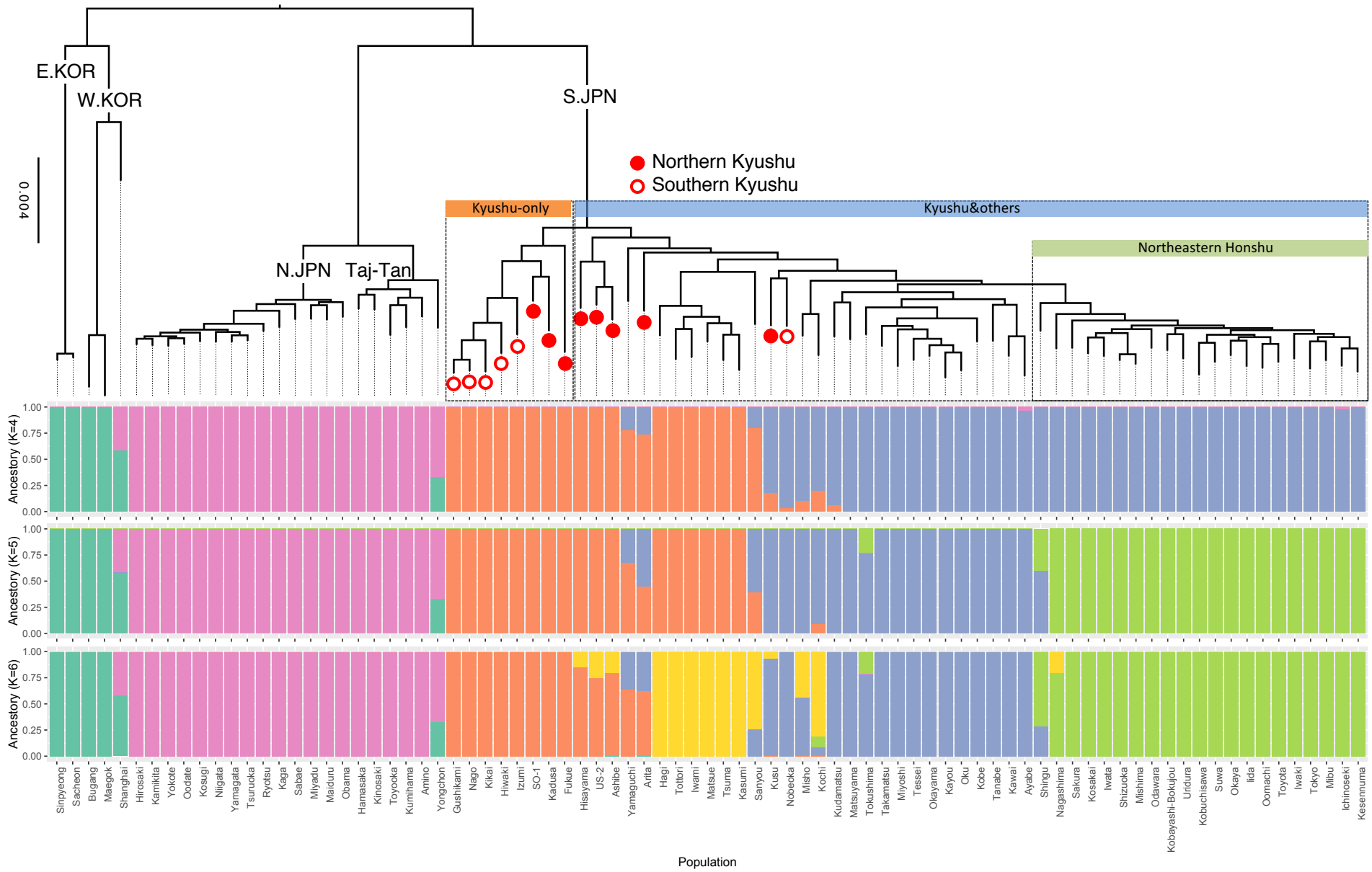
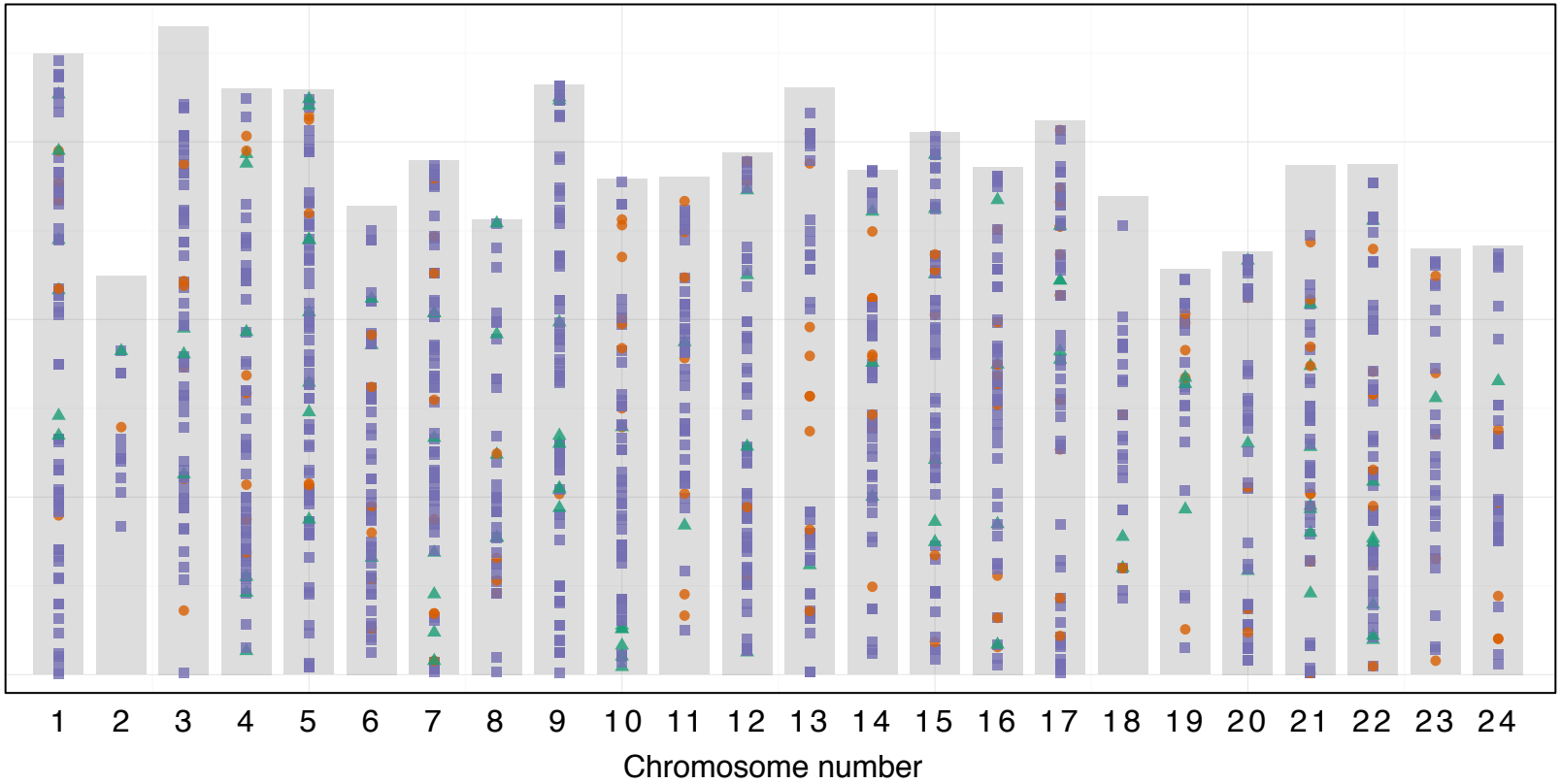

E.KOR
W.KOR
S.JPN
N.JPN
Taj-Tan
0.004
Northern Kyushu
Southern Kyushu
Kyushu-only
Kyushu&others
Northeastern Honshu
Ancestory (K=4)
Ancestory (K=5)
Ancestory (K=6)
1.00
0.75
0.50
0.25
0.00
Sinpyeong
Sacheon
Bugang
Maegok
Shanghai
Hirosaki
Kamikita
Yokote
Oodate
Kosugi
Niigata
Yamagata
Tsuruoka
Ryotsu
Kaga
Sabae
Miyadu
Maiduru
Obama
Hamasaka
Kinosaki
Toyooka
Kumihama
Amino
Yongchon
Gushikami
Nago
Kikai
Hiwaki
Izumi
SO-1
Kadusa
Fukue
Hisayama
US-2
Ashibe
Yamaguchi
Arita
Hagi
Tottori
Iwami
Matsue
Tsuma
Kasumi
Sanyou
Kusu
Nobeoka
Misho
Kochi
Kudamatsu
Matsuyama
Tokushima
Takamatsu
Miyoshi
Tessei
Okayama
Kayou
Oku
Kobe
Tanabe
Kawai
Ayabe
Shingu
Nagashima
Sakura
Kosakai
Iwata
Shizuoka
Mishima
Odawara
Kobayashi-Bokujou
Uridura
Kobuchisawa
Suwa
Okaya
Iida
Oomachi
Toyota
Iwaki
Tokyo
Mibu
Ichinoseki
Kesennuma
Population

Figure 4



	Mark	Chr1	Chr2	Chr3	Chr4	Chr5	Chr6	Chr7	Chr8	Chr9	Chr10	Chr11	Chr12	Total*1
NJPN–Tajima-Tango	■	62	11	52	54	70	53	65	32	75	48	44	61	627
SJPN–Tajima-Tango	●	7	1	6	7	6	6	8	5	1	8	7	4	66
Tajima-Tango only	▲	6	1	3	6	8	3	7	5	8	7	2	4	60

	Mark	Chr13	Chr14	Chr15	Chr16	Chr17	Chr18	Chr19	Chr20	Chr21	Chr22	Chr23	Chr24	Total*2
NJPN–Tajima-Tango	■	39	47	60	50	51	28	21	44	42	52	26	36	496
SJPN–Tajima-Tango	●	10	9	6	9	13	3	5	5	9	8	5	6	88
Tajima-Tango only	▲	1	3	6	4	5	2	3	3	7	7	1	1	43

Figure 5

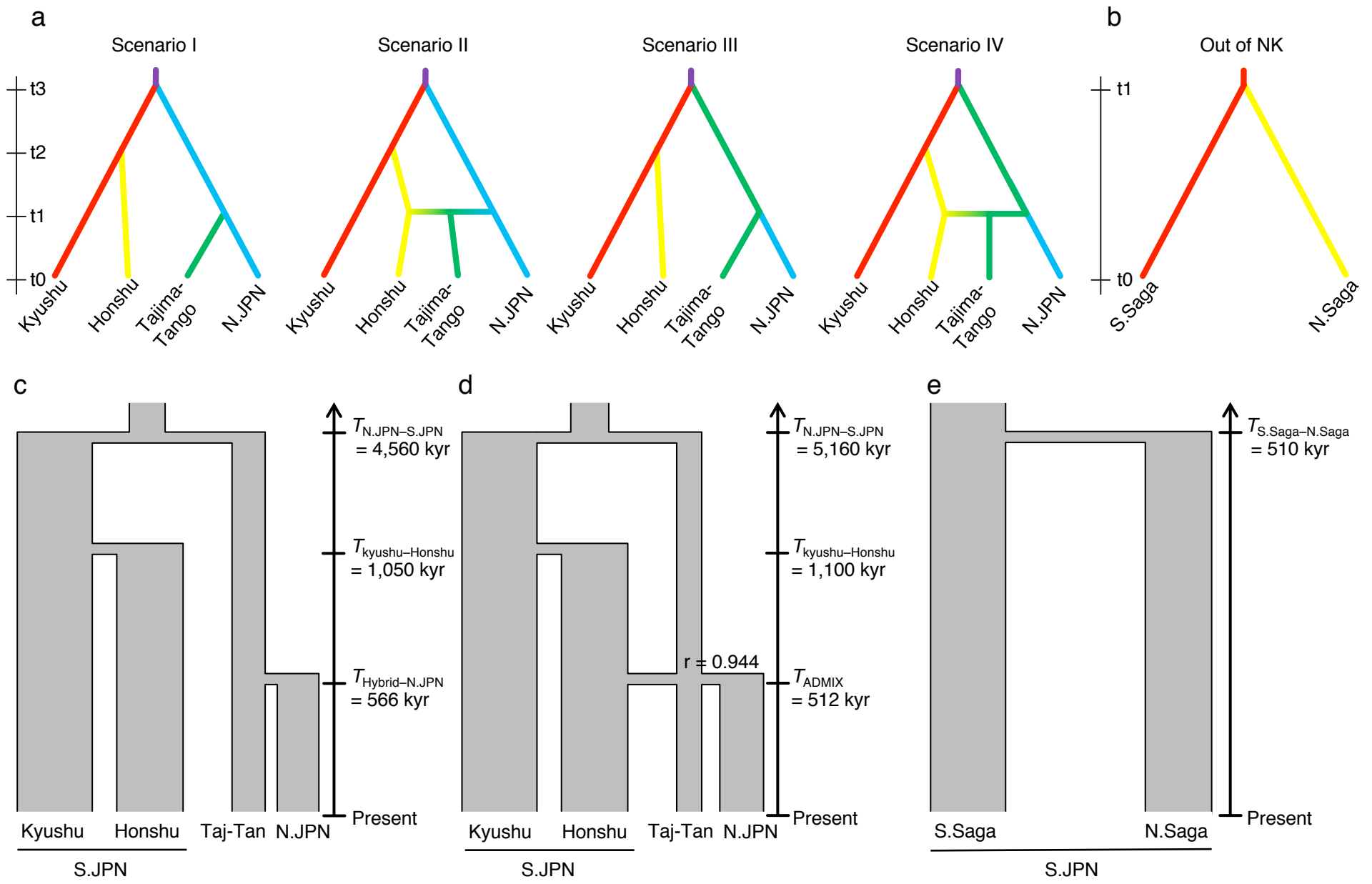
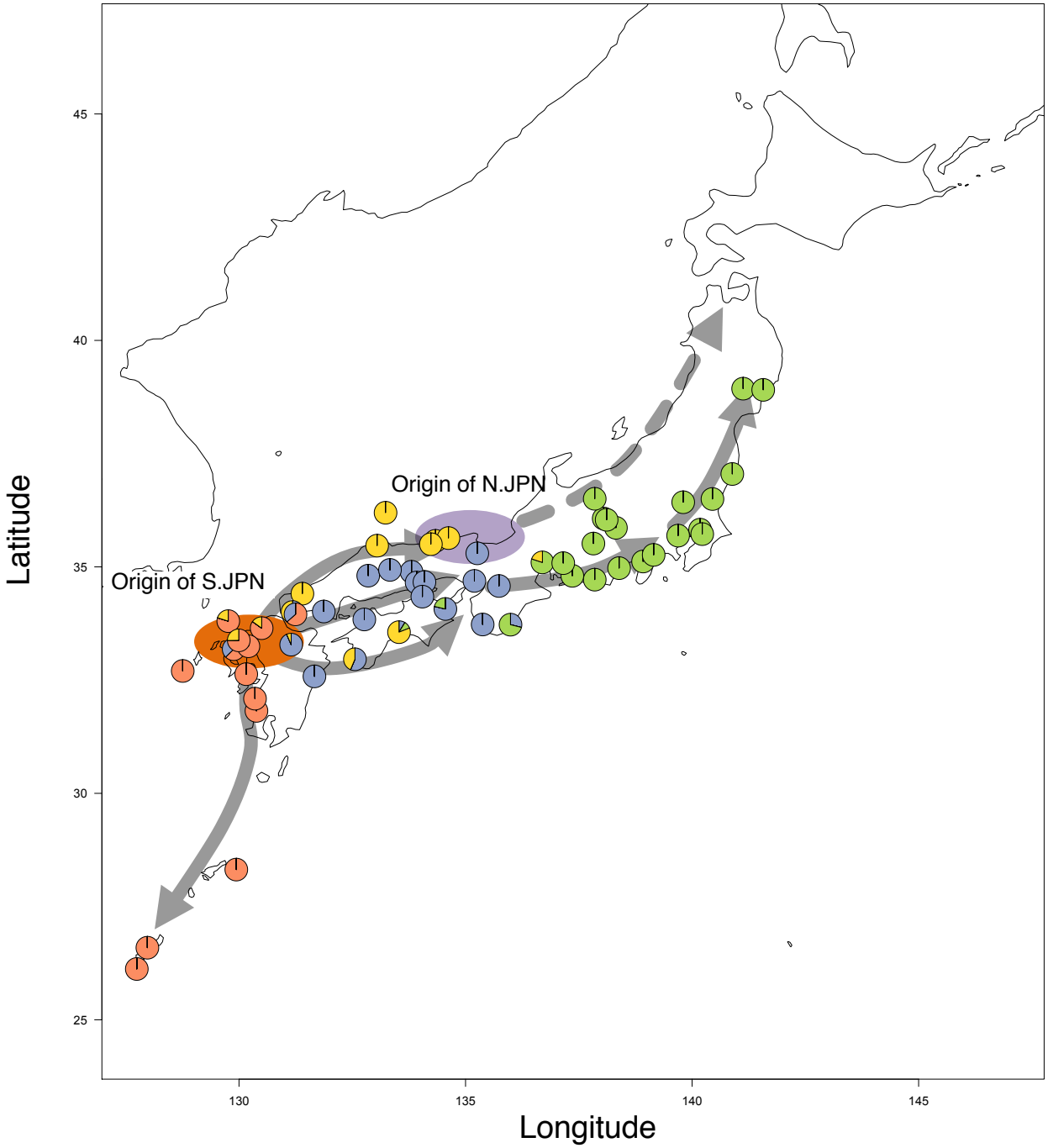

a
Scenario I
Scenario II
Scenario III
Scenario IV
t3
t2
t1
t0
Kyushu
Honshu
Tajima-Tango
N.JPN
b
Out of NK
t1
t0
S.Saga
N.Saga
c
$T_{N.JPN–S.JPN}$ = 4,560 kyr
$T_{kyushu–Honshu}$ = 1,050 kyr
$T_{Hybrid–N.JPN}$ = 566 kyr
Present
Kyushu
Honshu
Taj-Tan
N.JPN
S.JPN
d
$T_{N.JPN–S.JPN}$ = 5,160 kyr
$T_{kyushu–Honshu}$ = 1,100 kyr
r = 0.944
T_{ADMIX} = 512 kyr
Present
Kyushu
Honshu
Taj-Tan
N.JPN
S.JPN
e
$T_{S.Saga–N.Saga}$ = 510 kyr
Present
S.Saga
N.Saga
S.JPN

Figure 6


Origin of N.JPN
Origin of S.JPN
Latitude
Longitude
45
40
35
30
25
130
135
140
145