ICES Journal of
Marine Science

ICES CIEM
International Council for the Exploration of the Sea
Conseil International pour l'Exploration de la Mer

ICES Journal of Marine Science (2016), 73(9), 2181–2189. doi:10.1093/icesjms/fsw070

Editor's Choice
Reproductive skew in Japanese sardine inferred from DNA sequences

Hiro-Sato Niwa*, Kazuya Nashida, and Takashi Yanagimoto

National Research Institute of Fisheries Science, Yokohama 236-8648, Japan

**Corresponding author: tel: +81 45 788 7691; fax: +81 45 788 5001; e-mail: hiro.s.niwa@affrc.go.jp*

Niwa, H-S., Nashida, K., and Yanagimoto, T. Reproductive skew in Japanese sardine inferred from DNA sequences. – ICES Journal of Marine Science, 73: 2181–2189.

Received 6 November 2015; accepted 5 April 2016; advance access publication 18 May 2016.

An excess of low-frequency mutations is a ubiquitous characteristic of many marine species, and may be explained by three hypotheses. First, the demographic expansion hypothesis postulates that many species experienced a post-glacial expansion following a Pleistocene population bottleneck. The second invokes some form of natural selection, such as directional selection and selective sweeps. The third explanation, the reproductive skew hypothesis, postulates that high variation in individual reproductive success in many marine species influences genetic diversity. In this study, we focused on demography and reproductive success and the use of coalescent theory to analyse mitochondrial DNA sequences from the Japanese sardine. Our results show that population parameters estimated from both the site-frequency spectrum and the mismatch distribution of pairwise nucleotide differences refute the demographic expansion hypothesis. Further, the observed mismatch distribution, compared with the expectations of the reproductive skew hypothesis, supports the presence of multiple mergers in the genealogy. Many short external branches but few long terminal branches are found in the sardine genealogy. Model misspecification can lead to misleading contemporary and historical estimates of the genetically effective population sizes in marine species. The prevalence of reproductive skew in marine species influences not only the analysis of genetic data but also has ecological implications for understanding variation in reproductive and recruitment patterns in exploited species.

Keywords: coalescent, effective population size, excess singletons, multiple mergers, recruitment variation, skewed offspring-number distribution.

Introduction

A growing body of molecular data show that many marine species with broadcast spawning have star-like gene genealogies. Many studies have concluded from the analysis of mitochondrial DNA (mtDNA) sequences with coalescence-based methods, such as Bayesian skyline plots (BSPs), or the analysis of nucleotide differences between sequences, that populations of many marine species experienced a recent rapid demographic expansion (e.g. Crandall *et al.*, 2012). Under this interpretation, populations were depressed by the environmental effects of the Last Glacial Maximum, and post-glacial warming promoted population expansions. Looking back in time, a bottleneck generates a brief period of rapid coalescences in a gene genealogy just before a population expansion, and mutations appearing after the expansion occur on external branches of the genealogy and appear as singleton haplotypes in a sample. As a consequence, sequence mismatch distributions for these populations are unimodal (Slatkin and Hudson, 1991; Rogers and Harpending, 1992). It is noteworthy that the putative onset of population expansion in many marine species predated the last glacial maximum 18–20K years ago (Janko *et al.*, 2007; Grant, 2015; Sromek *et al.*, 2015), suggesting that populations were unaffected by subsequent glacial cycles.

The multiple-merger model based on reproductive skew among individuals (e.g. Eldon, 2011) is an alternative to the population expansion model and can account for the excesses of singleton mutations (Durrett and Schweinsberg, 2005; Berestycki *et al.*, 2007). The reproductive skew model is based on the life history trait common to most marine fish and invertebrates with type-III survivorship curves, which spawn large numbers of eggs producing larvae with high mortality rates early in life. The heavy-tailed, power-law distribution has been used to model the distribution of offspring number among individuals (Schweinsberg, 2003). When highly successful reproduction events occur, a large fraction of the population arises from only a few individuals who win 'reproduction sweepstakes' (Beckenbach, 1994; Hedgecock, 1994; Hedgecock and Pudovkin, 2011). Marine fish and invertebrate populations are also characterized by high variability in recruitment of young fish into a population. The statistical properties of the sum of power-law distributed random numbers of offspring then specify the variation in recruitment, as described by the generalized central limit theorem (Zaliapin *et al.*, 2005). Stochastic variation in recruitment, if mediated by a heavy tail in the distribution of reproduction success, should produce exceptionally strong year classes. The variance in recruitment is much greater than can be accommodated by reproduction models in the domain of attraction of the Gaussian stable law (Feller, 1971). Distinguishing between the effects of reproductive skew and population expansions on DNA sequences can provide information about the large fluctuations in recruitment.

The neutral ‘Kingman’ coalescent model (Kingman, 1982; Hudson, 1983; Tajima, 1983) is the predominant model used to generate null hypotheses for interpreting genetic data. In simulations of gene genealogies, the Kingman coalescent allows for only pairwise mergers of ancestral lines, whereas reproductive skew models can accommodate multiple mergers (Schweinsberg, 2003). The constructions of coalescent processes with multiple mergers were first studied independently by Donnelly and Kurtz (1999), Pitman (1999), and Sagitov (1999). The gene genealogies produced by these two coalescent models can be quite different. When sweepstakes recruitment events occur, the majority of the individuals in a sample come from only a few individuals and this produces multiple-merger coalescences in a gene genealogy. The effect of multiple-merger coalescence in inferring population characteristics has been studied theoretically in recent years (Eldon and Wakeley, 2006; Sargsyan and Wakeley, 2008; Eldon, 2011; Steinrücken *et al.*, 2013). However, these theoretical models have only rarely been used to understand mtDNA sequence diversity in marine species. One exception is the study by Eldon *et al.* (2015), who investigated the fits of the multiple-merger coalescent model and the Kingman coalescent (with exponential population-growth) model to Atlantic cod mtDNA (cytochrome *b*) sequence data. They concluded, however, that the number of segregating sites was not large enough to distinguish between the two models. Another study by Árnason and Halldórsdóttir (2015) showed that both the population-growth model and the multiple-merger coalescent model equally captured the high frequency of singleton sites for nuclear genes (haemoglobin A_2 and myoglobin) in the Atlantic cod. However, this comparison did not constitute a formal test. Instances of multiple-merger coalescents may be common, but have not been recognized.

The objective of our study was to test whether the control region (CR) of mtDNA for Japanese sardine *Sardinops melanostictus* reflected a recent population expansion, or has been imprinted by reproductive skew. We approximated the genealogy of the sample by coalescent processes under the Kingman coalescent and multiple-merger models and compared the fits of these models with the characteristics of the sardine sequences. The results indicate that a population expansion is unlikely and that sweepstakes reproductive success is a more likely explanation for the pattern of sequence variation in the Japanese sardine. The results also provide an insight into the importance of individual-level stochasticity on population-level outcomes even in abundant species.

Materials

Specimens of sardine were collected during the spawning season (February and March) in 1990 from Tosa Bay, Japan, and were frozen at −20°C until DNA extraction. Genomic DNA was isolated from muscle tissue by QuickGene-810 (Toyobo, Osaka, Japan). We used the complete sequence of *Sardinops melanostictus* mtDNA (NC002616) to design the SMThrL primer, 5′-tgccccagtagctta gttcaa-3′ (forward), and the SMPheH primer, 3′-gcttcttacggccca tctta-5′ (reverse) to amplify the mtDNA CR. Each PCR was performed in 10 mm^3 containing 5–10 ng template DNA, 1 mm^3 of 10× reaction buffer, 1 mol m^{-3} each dNTPs, 2×10^{-3} mol m^{-3} each primer, and 0.5 units of EX Taq HS polymerase (Takara, Shiga, Japan) in an ABI 9700 Thermal cycler (Applied Biosystems, Foster City, CA, USA). Initial denaturation was for 2 min at 94°C; followed by 40 cycles of 30 s at 94°C for denaturation, 30 s at 55°C for annealing, and 30 s at 72°C for extension; and a final extension at 72°C for 10 min. PCR product was purified with GFX (Qiagen, Hilden, Germany) and was used as the template DNA for cycle sequencing reactions performed using Big Dye Terminator Cycle Sequencing Kit (Ver.3.1, Applied Biosystems). Sequencing was conducted on an ABI Prism 3730XL (Applied Biosystems) automatic sequencer with four primers containing PCR primers and inner primers (SMCRL: 5′-gtagtaagaaccgaccaaccg-3′ and PheHR primer: 5′-tcaaagcataacactgaagatgttaa-3′). The mtDNA CR was 1212 bp long for 106 sardine. These sequences were deposited in GenBank (accession nos. LC031518–LC031623).

Methods and results

Preliminary exploration of mtDNA sequence variation

We identified 101 haplotypes among the 106 sardine sequences. Most haplotypes were in low frequency: 98 were observed only once, 2 were observed twice, and 1 was observed four times. Figure 1 shows the folded site-frequency spectrum (SFS) for the sardine mtDNA CR sequences. Compared with the expected SFS in a neutral equilibrium population assuming the Kingman coalescent under an infinitely-many-sites model (ISM) of mutation (Watterson, 1975; Tajima, 1989), we observed an excess of singleton sites. An abundance of low-frequency mutations likely resulted from a star-like topology with a short genealogy, whereas middle-frequency polymorphisms would be a signature of long population history. An unrooted maximum likelihood tree was constructed from the mtDNA CR sequences using program `phyml` (Guindon *et al.*, 2010). The tree branched unevenly (Figure 2). There was a burst of mergers at one vertex, and few edges of the tree merged together at the other vertices. There were mergers at different depths in the tree.

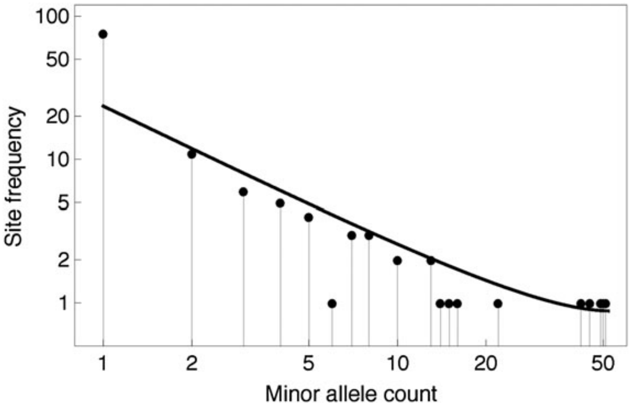

Figure 1. Folded SFS for the entire mtDNA CR sequences of sardine. The solid line is the expected SFS under constant population size, assuming the ISM given the number 122 of segregating sites in a sample of size 106.

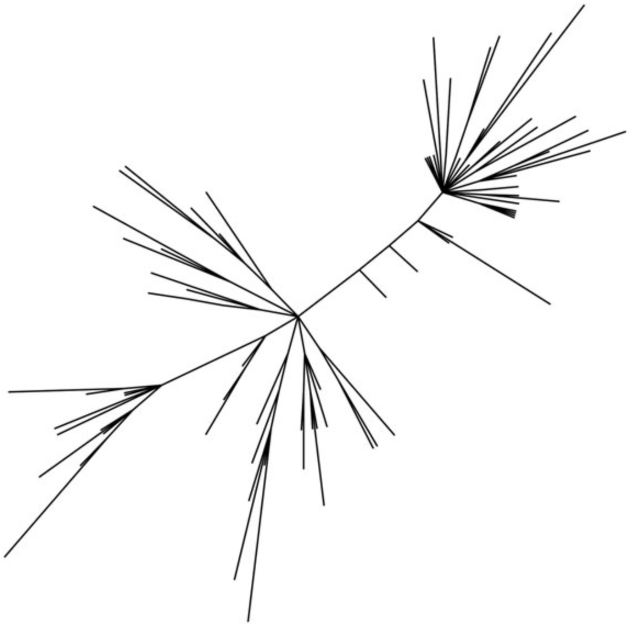

Figure 2. Unrooted maximum likelihood tree in PhyML (Guindon *et al.,* 2010), constructed from the sardine mtDNA CR sequences without invariable sites. K80+Γ was the one that best explained the data in terms of AIC. Forty-four edges collapsed to the upper right vertex.

Infinitely-many-sites model

Computational models of demographic inference in this study assume the ISM, in which every mutation occurs at a unique site as a Poisson process on the edges of the ancestral tree. There were 122 segregating sites in the entire mtDNA CR (1212 bp) of sardine, and some pairs of the sites cannot be fitted on a single gene tree topology (Felsenstein 1982). We therefore selected a subset of sites to which the ISM applied. We solved the violations of the ISM by excluding topologically incompatible sites in an unrooted sense using program `genetree` (Griffiths, 2002) and found the largest set of sites that are mutually compatible, resulting in a total of 80 segregating sites defining 51 haplotypes (see Supplementary Appendix A and Figure S1). Our choice of the ISM is based on the belief that, conditional on the ancestral tree of a subset of sites compatible with the ISM, mutations occur as a Poisson process. This claim follows from the thinning property of the Poisson process (Kingman, 1993). Therefore, this treatment is not likely to bias the analyses.

Sardine sequence data exhibit recurrent mutations at a substantial number of sites, which may be common in the mtDNA CR for a variety of species. For example, there were 75 segregating sites in 71 sequences (partial CR, 395 bp) from Pacific bluefin tuna *Thunnus orientalis* (Nomura *et al.*, 2014), and we found that 29 segregating sites (22 haplotypes) were compatible with the ISM. For another example, there were 90 segregating sites in 51 sequences (partial CR, 615 bp) from Japanese eel *Anguilla japonica* (Ishikawa *et al.*, 2001), and we found that 60 segregating sites (39 haplotypes) were compatible with the ISM. Note that a high fraction of recurrent mutation sites would result in the underestimation of the substitution rate for the mtDNA CR.

The ISM assumption implies that an unrooted tree can be constructed uniquely from a collection of data sequences (Griffiths, 2002). Figure 3 shows the unrooted gene tree constructed from the sardine mtDNA CR sequences compatible with the ISM. All possible rooted trees may be found from an unrooted tree by placing the root at a vertex or between mutations. We used a likelihood method under the Kingman coalescent framework (Griffiths, 2002) for estimating the position of the root in the unrooted tree (see Supplementary Appendix B). There are 81 possible rooted trees, and with probability 0.625 the most likely root is placed between the mutations labelled 12 and 13 (hereafter referred to as the ML rooted tree). The neutral model predicts that the most frequent allele at each site is likely to be the oldest (Watterson and Guess 1977). If we hypothesized the root based on frequencies, it would include sequences labelled Ω with frequency 54/106, and the root is placed at sequence type Ω with probability 6.74×10^{-3}.

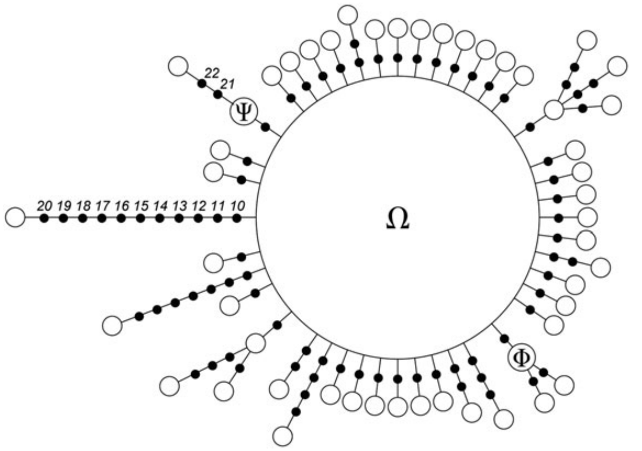

Figure 3. Unrooted gene tree constructed from the sardine mtDNA CR sequences (compatible with the ISM). Open circles represent sequence types and bold dots along the edges represent mutations. The sequence type labelled Ω occurs in the sample with frequency 54 of 106 sequences, and sequences Ψ and Φ both occur twice. The unlabelled vertices correspond to sequence types with frequency 1/106.

Bayesian skyline inference of population size history

We used a Bayesian Markov chain Monte Carlo (MCMC) method under the Kingman coalescent framework implemented in the BEAST 2 (Drummond and Bouckaert, 2015) to construct a coalescent tree from the sardine mtDNA CR sequences. We ran the MCMC routine for 2×10^9 iterations with a burn-in of 10% and with sampling every 50 000. This yielded 40 001 rooted trees. All parameters had effective sample sizes ≥488.

We calculated the rooted tree with the highest product of clade posterior probabilities. The maximum clade credibility (MCC) tree (Figure 4a) was characterized by a topology with uniformly long terminal branches. All coalescent events occurred near the root. An excess of external branches in the coalescent tree resulted in an excess of singleton mutations. The mutations that arose on this tree have likely been inherited by only a single DNA copy in the sample, and thus, the most frequent allele at a nucleotide site is likely to be ancestral, which holds under the Kingman coalescent (Sargsyan and Wakeley, 2008). Figure 4b shows a BSP of population sizes. The ordinate is the product of the effective population size (of females) and 2μ, where μ is the aggregate mutation rate per generation. The abscissa (time before present) is measured in mutational time units corresponding to $(2\ \mu)^{-1}$ generations, where one unit

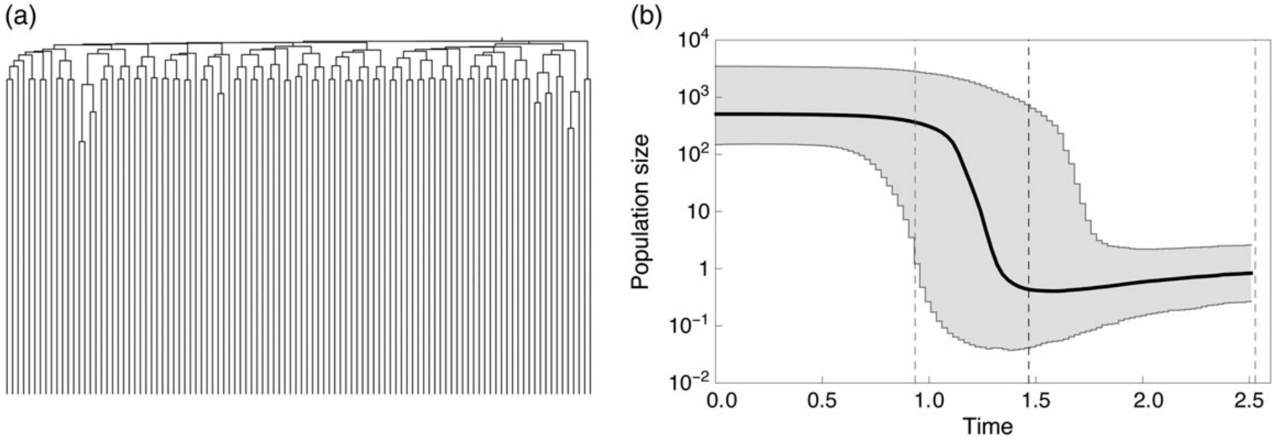

Figure 4. Bayesian coalescent analysis of the sardine mtDNA sequences, using the BEAST 2 software package (Drummond and Bouckaert, 2015). (a) MCC coalescent tree. This tree corresponds to the most frequent allele at each site being labelled as ancestral. (b) BSP. Shown are the median population size in the solid line (resp. median T_{MRCA} in vertical black dashed line) and the 95% highest posterior density interval in the shaded grey (resp. enclosed by grey dashed lines).

represents the time needed for a single mutational difference to accumulate between two lineages. The median time to the most recent common ancestor (MRCA) of the sample, the median root height of the coalescent tree, was $T_{\mathrm{MRCA}} = 1.467$. The inferred median values of population size N_0 at present (time of sampling) and N_1 at median T_{MRCA} are θ_0 $(=2N_0\mu) = 506$ and θ_1 $(=2N_1\mu) = 0.434$ in units of $(2\ \mu)^{-1}$ genes. A possible interpretation of this plot is that the population underwent a rapid expansion of three orders of magnitude. Japanese sardine appear to have experienced two epochs of population size over recent evolutionary history.

Population expansion model

We considered the two-epoch demographic model, in which the population instantaneously changed at time τ from an ancestral size N_1 to a current size N_0. The ancestry of a sample was approximated by a Kingman coalescent. Demographic history is scaled in time units of $(2\ \mu)^{-1}$ generations. Coalescent events occur at rate $1/\theta_0 = (2N_0\mu)^{-1}$ in the post-expansion period, and the expected pairwise nucleotide difference in an ancestral, equilibrium population is $\theta_1 = 2N_1\mu$. We used the Poisson random field (PRF) framework (Sawyer and Hartl, 1992) to infer the demographic-history parameters from the SFS of the sardine mtDNA sequences (see Supplementary Appendix C). The maximum likelihood estimates (MLEs), found using program `prfreq` (Boyko *et al.*, 2008), were identical in both the ML and MCC rooted trees: $\theta_0 = 1000$ (95% confidence interval: 802–1243), $\theta_1 = 0.89$ (0.828–2.15), and $\tau = 1.5$ (1.17–1.55). The ratio of ancestral to current effective population size, $N_1/N_0 = 8.9 \times 10^{-4}$ (95% CI: 8.28×10^{-4}–21.5×10^{-4}). From 10^4 replications of trees simulated with the program `ms` (Hudson, 2002) with MLEs, the median root height was $T_{\mathrm{MRCA}} = 3.044$ with a 2.5–97.5% quantile-interval of 2.102–5.912. The estimated parameters were consistent with those estimated from the BSP. The MLE under the standard Kingman coalescent model ($\tau = \infty$ and $N_1/N_0 = 1$) was equal to Watterson's (1975) estimator for the sample, $\theta_W = 15.3$. The likelihood ratio test (LRT) statistic is 220.4. The null hypothesis of equilibrium was therefore rejected.

Reproductive skew model

We now consider coalescents arising from the reproductive skew model in a population with constant size N. The number of offspring produced by each individual is a random variable with mean $m > 1$. The probability of having x or more offspring is assumed to follow a heavy-tailed, power-law distribution $Cx^{-\alpha}$ with some constants $C > 0$ and $1 < \alpha < 2$. The variance of the offspring number is unbounded as $N \to \infty$. The skew becomes stronger for smaller values of α. The total number of recruits entering the population will be typically much larger than N, but only some offspring will survive to the next generation. In each generation, N offspring are chosen at random to reproduce. In such a genealogical model for populations with large offspring sizes, arbitrary multiple mergers are possible (Schweinsberg, 2003). The probability that two or more lineages sampled uniformly from the current generation derive from a common ancestor in the immediately previous generation depends only on the number of lineages involved in a merger. The coalescence probability, which is the probability that two individuals chosen at random out of N share the same common ancestor, is given by

$$c_N = C\alpha\ m^{-\alpha}\mathrm{Beta}(2-\alpha,\alpha)N^{1-\alpha}, \qquad (1)$$

in the large N limit (Schweinsberg, 2003), where $\mathrm{Beta}(a,\ b) = \Gamma(a)\Gamma(b)/\Gamma(a+b)$. Then the underlying 'Beta(2−α,α)' coalescent involves a power-law time scaling according to $N^{\alpha-1}$; that is, mutations appear on coalescent branches at the Poisson rate $\theta_B/2$ in coalescent time units of $1/c_N$ generations. The scaled mutation rate is defined by $\theta_B = 2\ \mu/c_N$. Note that the range $\alpha \geq 2$ gives rise to the standard Kingman coalescent (Schweinsberg, 2003), and the variance of the offspring number has a finite limit as $N \to \infty$. In Supplementary Appendix D, we provide a concise introduction to the main statistical features of Beta coalescents, and a link with generalized central limit theorem for stable random variables.

We considered the Beta(2−α,α) coalescent as the underlying model, which yielded a coalescent history that was consistent with the sardine mtDNA gene tree. We computed the likelihood of observed gene tree under the parameters ($\theta_B/2$, α) employing an importance sampling scheme in the ISM using program `MetaGeneTree` (Birkner *et al.*, 2011). Likelihood values were estimated independently for each discrete gridpoint using 10^6 independent runs of the Markov chain with spacing $(\Delta\theta_B/2,\ \Delta\alpha) = (0.25,\ 0.02)$ between gridpoints.

The MLEs were $\theta_B/2 = 3.00$ and $\alpha = 1.28$ for the ML rooted tree, and $\theta_B/2 = 3.25$ and $\alpha = 1.30$ for the MCC tree, where the log likelihoods were 16.80 and 17.81, respectively. Both estimates were not statistically different (Figure 5 and Supplementary Figure S3). The likelihood function under the null hypothesis of standard Kingman coalescent ($\alpha = 2$) was numerically evaluated for both trees and had the value of zero, $L(1 \leq \theta_B/2 \leq 20, \alpha = 2) = 0$. We, therefore, can refute the standard Kingman coalescent under the reproduction model in the domain of attraction of the Gaussian stable law. The value of the coalescent parameter $\alpha < 2$ was significant. From 10^6 replications of trees simulated with the program `MetaGeneTree` with MLEs for the ML rooted tree (resp. MCC tree), the median root height was $T_{\mathrm{MRCA}} = 11.78$ (resp. 8.00) with a 2.5–97.5% quantile-interval of 6.94–17.29 (resp. 4.13–10.47) in mutational time units. Estimates under the Beta coalescent are even larger than T_{MRCA} under the population expansion model.

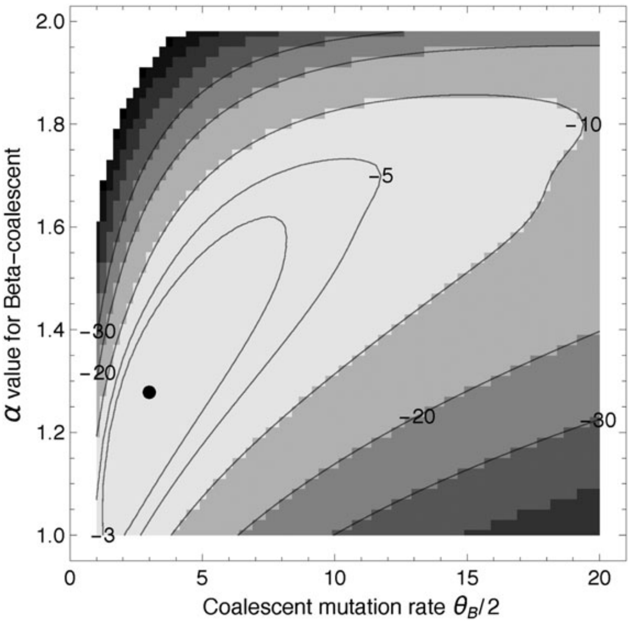

Figure 5. Log-likelihood surface (scaled to maximum likelihood) for the ML rooted gene tree of sardine mtDNA sequences analysed under the Beta($2-\alpha$,α) coalescent, outputted by program `MetaGeneTree` (Birkner *et al.*, 2011). The arg-maximum of the likelihood surface is indicated by a dot. The 95% joint confidence contour (likelihood based) is defined by taking the values of ($\theta_B/2$, α) for which the natural logarithm of the likelihood is $\chi^2_{0.95}(2)/2 = 2.995$ units smaller than the log of maximum likelihood, where $\chi^2_{0.95}(2)$ is the 0.95 quantile of a χ^2-distribution with two degrees of freedom.

Self-consistency check

Making pseudo-samples

We generated replicate samples from the population expansion model with SFS-based MLEs using the program `ms`, in which an ancestral tree by random backward coalescences is first generated then, mutations are placed at the points of Poisson processes, independently on each branch of the tree. Random samples from the reproductive skew model were generated by the time-reversed block counting process (Birkner and Blath, 2008). Both algorithms generate samples under the ISM. A total of 5000 replications of 106 sequences were simulated for each of the two models. The model is considered to predict the observed data well, if the observed site-frequency counts and pairwise difference counts fall within the 2.5–97.5% quantile-intervals obtained from the simulations.

Site-frequency spectra

Replicates under the two models were similar at low frequencies, but showed noticeable differences in the right tails of the SFS (Figure 6). The simulated SFS dominated at low frequencies in both of the models. A characteristic upturn at high derived frequencies appeared in the simulated SFS under the Beta($2-\alpha$,α) coalescent, which stems from old splits in the coalescent time. At branching events deep in the tree, almost all individuals of the population descended from one branch and inherited the mutations arising on that branch, whereas few descended from the other branches. Consequently, the SFS of derived alleles has peaks at near fixation. Such skewed branching is unlikely in the population expansion model under the Kingman coalescent. A peak near fixation was observed in the SFS of the ML rooted tree for the sardine mtDNA sequences,

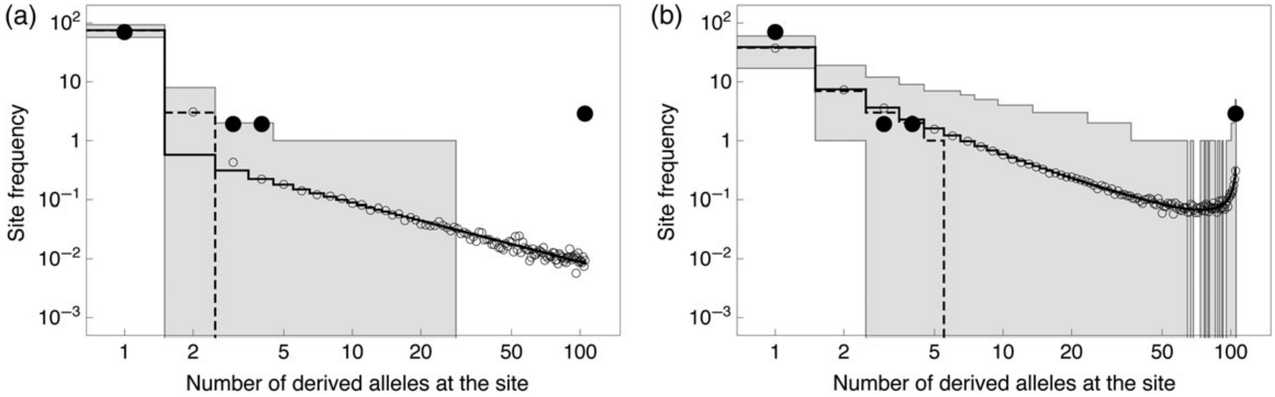

Figure 6. SFS of pseudo-samples of size 106 made by coalescent simulations under the population expansion model (a) and the reproductive skew model with MLEs for the ML rooted tree (b). The shaded area represents the 2.5–97.5% quantile-intervals for each class. The bold dots show the unfolded SFS (given the ML rooted tree) of sardine mtDNA sequences. The dashed staircase lines (resp. open circles) depict the median (resp. mean) over pseudo-samples. The solid staircase line in (a) is the PRF prediction of the transient distribution of the SFS, found using program `prfreq` (Boyko *et al.*, 2008). The solid staircase line in (b) shows the expected SFS associated with the Beta($2-\alpha$,α) coalescent at $\alpha = 1.28$, computed with the program `cmpfrlambda` (Birkner *et al.*, 2013), where mutations were generated at rate $\theta_B = 6.00$.

whereas such a peak was absent in the SFS given the MCC tree (Supplementary Figure S4a). Hence, the SFS alone would not provide enough information to distinguish between the population expansion and reproductive skew models, if we were not sure about the position of the root in the sardine mtDNA CR tree.

Note that the SFS under the Beta($2-\alpha,\alpha$) coalescent at low frequencies i (much less than the sample size) is proportional to $i^{\alpha-3}$ (Berestycki *et al.*, 2007) and hence is much steeper than the neutral SFS ($\sim i^{-1}$) under the standard Kingman coalescent (Tajima, 1989). The Beta coalescent tends to have more rare alleles than the standard neutral coalescent.

Pairwise mismatch distribution analysis

The distribution of pairwise differences is independent of the root position in the gene tree. The observed distribution of pairwise differences in the sardine mtDNA sequences (Figure 7) showed a large fraction of the pairwise comparisons with zero differences ($s = 0$), which means that there are many short external branches in the coalescent tree. Figure 4a shows that coalescences are concentrated in a narrow interval of the expansion and thus, correlations between pairwise coalescence times imposed by their common history given the MCC tree are unimportant (Slatkin and Hudson, 1991). Assuming that each pairwise comparison was independent among mtDNA lineages with a history of demographic expansion, we obtained the MLEs of population expansion parameters based on the mismatch distribution (Rogers and Harpending, 1992): $\theta_0 = \infty$ (95% CI: $62.5-\infty$), $\theta_1 = 1.25$ (1.20–1.29), and $\tau = 0.45$ (0.404–0.475); see Supplementary Appendix E for our pairwise differences based estimation method. The MLE under the standard Kingman coalescent model ($\theta_0 = \theta_1$ and $\tau = \infty$) was equal to Tajima's (1983) estimator $\theta_T = 1.69$, which is the average number of pairwise difference in the sample. The LRT statistic is 397.8. The null hypothesis of equilibrium was therefore rejected.

Applying the Mathematica non-linear fitting routine (Ver.10.4, Wolfram Research, Champaign, IL, USA) to the mismatch distribution for each of the 5000 pseudo-samples under the population expansion model yielded 2.5–97.5% quantile-intervals for the parameters: $159-9.72 \times 10^7$ for θ_0, $2.96 \times 10^{-3}-3.88$ for θ_1, and 0.617–3.36 for τ. These values are consistent with the estimates from the SFS-based analysis of the sardine mtDNA sequences. The distributions of pairwise differences for the pseudo-samples generated under the population expansion model with SFS-based MLEs are concordant with the expected distribution (Figure 7a). Therefore, the parameter estimates of demographic history based on the SFS and on the mismatch distribution agree, where the underlying genealogy is governed by the Kingman coalescent with a recent population expansion.

Reproductive skew vs. population expansion

We now compare the alternative hypotheses to explain the excess of low-frequency polymorphisms. The estimates of parameters from the sardine mtDNA sequences using the mismatch distribution analysis are inconsistent with the estimates based on the SFS analysis. The upper limit of the growth parameter θ_0/θ_1 was indeterminate. The estimate of the timing of population expansion from the mismatch distribution was much more recent than that estimated from the SFS. The confidence interval does not overlap with that based on the SFS. The conclusion is inescapable that the demographic history imprinted in the original mtDNA sequences differs from the history imprinted in the pseudo-samples. This difference implies that a past demographic expansion did not occur in Japanese sardine populations.

We performed a further consistency check. While the population expansion model predicted that the mode of pairwise differences is at $s = 1$, the observed mismatch counts in tail region were outside the 2.5–97.5% quantile-intervals obtained with the simulations (Figure 7a). Mismatch counts $s > 11$ under the population expansion model are unlikely, regardless of the root position in the sardine mtDNA gene tree. A significant increase in the frequencies with mismatch differences of $s > 11$, or with pairwise coalescence times of $\tau > 11$ in mutational time units, reflects deep divergence between these haplotype pairs. These mismatch pairs may be separated by heights significantly greater than the root height of the coalescent tree estimated based on the SFS. On the other hand, while the pseudo-samples from the reproductive skew model did not show a pronounced wave in the mismatch distribution, the observed distribution counts fell within the 2.5–97.5% quantile-intervals obtained with the simulations (Figure 7b and Supplementary

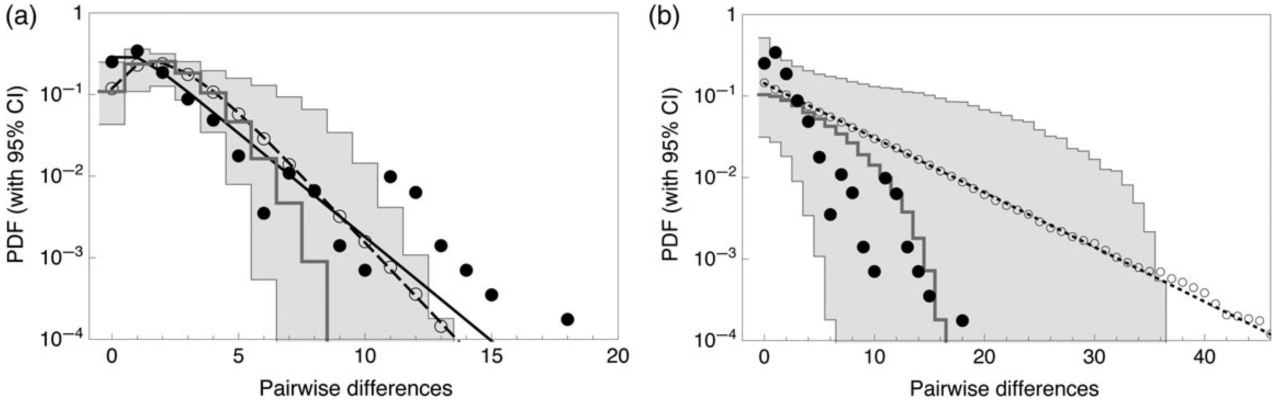

Figure 7. Mismatch distributions of pseudo-samples of size 106 made by coalescent simulations under the population expansion model (a) and the reproductive skew model with MLEs ($\alpha = 1.28$ and $\theta_B = 6.00$) for the ML rooted tree (b). The shaded area represents the middle 95% of the distribution of each mismatch count among the pseudo-samples. The bold dots show the pairwise differences observed in sardine mtDNA sequences. The solid line is the fitted curve using pairwise differences based MLEs. The thick grey staircase-lines (resp. open circles) depict the median (resp. the mean) over pseudo-samples. The dashed line (a) depicts the expected curve with SFS-based MLEs. The dotted line (b) shows the geometric distribution with parameter given by the value of θ_B (plus one) that was used to make the pseudo-samples.

Figure S4b). The reproductive skew model specifically predicted the tail region of the sample distribution, which is concordant with the median of the pseudo-samples.

Note that the expected mismatch counts for a population with reproductive skew converge to the geometric distribution (derivation in Supplementary Appendix D). The average pairwise difference under the reproductive skew model, i.e. the mean value of Tajima's estimator θ_T over the pseudo-samples (6.01 with 95% CI, 1.19–14.83 for the ML rooted tree, and 6.50 with 95% CI, 1.35–15.74 for the MCC tree) agrees with the value of θ_B that was used to generate mutations on the Beta coalescent branches. This follows from the coalescent time, or the time since the MRCA of two sequences, which is expected to be $1/c_N$, so that the expected difference in the number of mutations that have accumulated along the two lineages descended from the MRCA is θ_B.

Discussion

Reproductive skew models (or multiple-merger coalescent models) have recently been proposed as appropriate models to investigate highly fecund marine species. The empirical studies to distinguish between coalescents under the population expansion scenario and under the reproductive skew scenario are still largely missing. We contrasted the population expansion parameters estimated based on the SFS and the mismatch distribution, and showed that the sardine mtDNA sequence distribution refutes the population expansion scenario. We also compared the observed mismatch distribution with expectations in both the population expansion and reproductive skew models, and showed that the observed sequences support the presence of multiple mergers in the genealogy. There are many short external branches but few long terminal branches with deep coalescence times in the Japanese sardine tree. The tree branches unevenly, which is a basic deviation from the Kingman coalescent model, regardless of the history of population expansion. The Kingman coalescent assumption of pairwise mergers is not valid to analyse the mtDNA CR sequence variation in Japanese sardine. Notably, the BSP may not reflect demographic history in highly fecund marine species, and the MCC tree may represent a spurious phylogenetic relationship.

In stark contrast to the Kingman coalescent model for populations with relatively evenly distributed offspring, the multiple-merger coalescent model is anticipated by the skewed offspring-number distribution. Both the recent population expansion and reproductive skew models are able to generate excesses of low-frequency polymorphisms, as have been observed in mtDNA sequences for many marine populations. However, the two coalescent processes are conceptually distinct and produce gene genealogies that differ significantly in tree topology. Population bottlenecks and expansions lead to departures from drift-mutation equilibrium and can have long-lasting effects on genetic diversity. Allele frequencies have not changed rapidly in the post-expansion population, as expected from the Kingman coalescent. On the other hand, reproductive skew, which leads to multiple mergers in a gene genealogy, can alter allele frequencies significantly, even if population abundance remains high.

Although the theoretical aspects of coalescent processes with multiple mergers have been studied over the last decade (Eldon and Wakeley, 2006), only a few biological applications of the model have been made (Árnason and Halldórsdóttir, 2015). The choice of a population genetics model has ecological implications for understanding recruitment variation in marine species. If reproductive skew is a hallmark of abundant marine species such as sardine (*Clupeidae*) and cod (*Gadinae*), multiple-merger coalescent models can provide crucial insight into the variation in recruitment. We discuss this briefly below.

Effective population size

It will be important to investigate the effect of unevenness in reproductive success among individuals on the estimate of N_e, the effective population size. The variable N_e is a descriptor of allele frequency dynamics, defined as the size of a Wright-Fisher population in which the rate of change in allele frequencies is the same as in the studied population (Wright, 1931). The variable N_e can be estimated by the inverse of the probability that two randomly chosen lineages have the same parent in the previous generation (Sjödin *et al.*, 2005). The parameter θ_0 ($=2N_0\mu$), estimated from a single sample, provides an estimate of long-term N_e reflecting the entire coalescent history of a population under the expansion model. Sampled lineages rarely coalesce in the large post-expansion population and thus indicating that genetic drift is weak. This contrasts with the temporal methods in which the short-term N_e is estimated from the change of allele frequencies in two temporally spaced samples from different generations, or from the joint phylogeny of gene sequences from serial samples (Rodrigo and Felsenstein, 1999; Anderson, 2005).

Several studies show that in marine populations, estimates of short-term N_e ranges from the hundreds to the low thousands (Hauser and Carvalho, 2008). Thus, extremely low short- to long-term N_e ratios are expected in marine species. For example, genetic analyses of Mediterranean bluefin tuna *Thunnus thynnus* populations show that the long-term estimates are almost 100-fold larger than short-term estimates (Riccioni *et al.*, 2010). This may lead to the conclusion that genetic drift is strong in a population that temporarily decreases in size from overexploitation (Pinsky and Palumbi, 2014).

Our result suggests that the apparent reduction in N_e in a contemporary population is an artefact of the analysis due to not using multiple-merger coalescents in the analyses of samples. The effective population size of the multiple-merger coalescent model is given by $N_e = 1/c_N$ with the coalescence probability (Wakeley and Sargsyan, 2009; Huillet, 2014) and thus, $N_e = \theta_B/2\mu$ is expected to coincide with the temporal estimate of effective size. Thus, short-term N_e can be substantially smaller than long-term $N_e = \theta_0/2\mu$ predicted by the population expansion model. The ratio for Japanese sardine is small, $\theta_B/\theta_0 \approx 1/160$. A comparison of the short- vs. long-term N_e can therefore be used for a possible falsification of the population expansion scenario, because an extremely low short- to long-term N_e ratio clearly indicates the presence of multiple-merger gene genealogy due to reproductive skew.

In addition, N_e in the reproductive skew model can differ significantly from the census size N and does not necessarily vary linearly with N. To demonstrate this, let the offspring distribution be a heavy-tailed, power-law distribution on $x \geq C^{1/\alpha}$ with mean $m = C^{1/\alpha}\alpha/(\alpha-1)$. If $N_e = 10^3$, then Equation (1) with $\alpha = 1.28$ (resp. 1.3) gives $N = 1.93 \times 10^8$ (resp. 6.95×10^7) for the Japanese sardine. We can imagine that the loss of genetic variation through strong genetic drift is balanced by the gain through many new mutations each generation in a large population with high reproductive skew, where the coalescence rate c_N is large, when compared with that of the Kingman coalescent, because of multiple mergers in the gene genealogy. The effective size to census size ratio, $N_e/N \sim N^{\alpha-2}$, from Equation (1), decreases with increasing population size, which does actually occur in marine species (Hauser and

Carvalho, 2008). The biological mechanism for a power-law relation in the N_e/N ratio is clearly power-law skewed reproductive successes.

Recruitment variation

High-fecundity marine populations are characterized by intermittent, extremely large recruitments, which are critical for sustaining harvestable biomass (Leaf and Friedland, 2014). Standard fisheries models assume that recruitment comes from a lognormal distribution, which will occasionally lead to large recruitment. However, a test of this prediction is difficult because the detection of lognormal recruitment is unfortunately complicated by even larger fluctuations in recruitment (Hilborn and Walters, 1992). Time-series for fish stock abundances in the North Atlantic suggest that these outlying events occur far more frequently than expected from the lognormal projections (Niwa, 2006a, 2007). Since annual recruitment and spawning stock indices are in practice estimated from time-series of catch data, the variation in catch will have a similar behaviour. Annual domestically landed catch statistics, compiled by the Food and Agriculture Organization of the United Nations, confirmed this expectation (Niwa, 2006b).

The reproductive skew model assumes that each individual independently produces a random number of successful offspring following a power-law distribution with finite mean and infinite variance (i.e. unbounded as $N \to \infty$). If so, the generalized central limit theorem for stable random variables then holds (Zaliapin *et al.*, 2005; Huillet, 2014), so that recruitment, the sum of the independent and identically distributed numbers of offspring with a heavy-tailed, power-law distribution, follows a power-law distribution with the same tail index. Thus, the tail index provides information about the likelihood of rare events in recruitment. Our results suggest that the substantially large amount of variability in the tail of the recruitment distribution may be the rule rather than the exception. Time-series in ecology are generally too short to determine the tail index and thus, the sample variance of the recruitment time-series may be misleading for representing a measure of recruitment variation. We estimated that the tail index was $\alpha < 2$, by calibrating the Beta$(2-\alpha,\alpha)$ coalescent to the sardine mtDNA sequences. This value implies that, in years of extremely large recruitment, a rapid reduction in genetic variation occurs due to the disparity in reproductive success among individuals, as only a few spawners will win reproduction sweepstakes. Successful recruitments then occur concurrently with decreases in N_e. In the current paradigm of fisheries management of abundant species, recruitment is thought to be little influenced by among-individual variation in reproductive success. However, given the clear genetic signals of reproductive skew in marine species, understanding the extent of among individuals variation in reproductive success is essential for identifying the causes of variability in recruitment.

Supplementary data

Supplementary material is available at the *ICESJMS* online version of the manuscript.

Acknowledgements

We are very grateful for critical comments of anonymous reviewers that greatly improved the manuscript. We are also indebted to editor Stewart Grant for his encouragement. This work was partially supported by JSPS Grants-in-Aid for Scientific Research (25450292) and Project Study of NRIFS, Japan in 2012. In this research work, we used the supercomputer of AFFRIT, Japan.

References

Anderson, E. C. 2005. An efficient Monte Carlo method for estimating N_e from temporally spaced samples using a coalescent-based likelihood. Genetics, 170: 955–967.

Árnason, E., and Halldórsdóttir, K. 2015. Nucleotide variation and balancing selection at the *Ckma* gene in Atlantic cod: analysis with multiple merger coalescent models. PeerJ, 3: e786.

Beckenbach, A. T. 1994. Mitochondrial haplotype frequencies in oysters: neutral alternatives to selection models. *In* Non-Neutral Evolution, pp. 188–198. Ed. by B. Golding. Chapman & Hall, New York.

Berestycki, J., Berestycki, N., and Schweinsberg, J. 2007. Beta-coalescents and continuous stable random trees. Annals of Probability, 35: 1835–1887.

Birkner, M., and Blath, J. 2008. Computing likelihoods for coalescents with multiple collisions in the infinitely many sites model. Journal of Mathematical Biology, 57: 435–465.

Birkner, M., Blath, J., and Eldon, B. 2013. Statistical properties of the site-frequency spectrum associated with Λ-coalescents. Genetics, 195: 1037–1053

Birkner, M., Blath, J., and Steinrücken, M. 2011. Importance sampling for Lambda coalescents in the infinitely many sites model. Theoretical Population Biology, 79: 155–173.

Boyko, A. R., Williamson, S. H., Indap, A. R., Degenhardt, J. D., Hernandez, R. D., Lohmueller, K. E., Adams, M. D., *et al.* 2008. Assessing the evolutionary impact of amino acid mutations in the human genome. PLoS Genetics, 4: e1000083.

Crandall, E. D., Sbrocco, E. J., DeBoer, T. S., Barber, P. H., and Carpenter, K. E. 2012. Expansion dating: calibrating molecular clocks in marine species from expansions onto the Sunda Shelf following the Last Glacial Maximum. Molecular Biology and Evolution, 29: 707–719.

Donnelly, P., and Kurtz, T. G. 1999. Particle representations for measure-valued population models. Annals of Probability, 27: 166–205.

Drummond, A. J., and Bouckaert, R. R. 2015. Bayesian Evolutionary Analysis with BEAST. Cambridge University Press, Cambridge. 249 pp.

Durrett, R., and Schweinsberg, J. 2005. A coalescent model for the effect of advantageous mutations on the genealogy of a population. Stochastic Processes and Their Applications, 115: 1628–1657.

Eldon, B. 2011. Estimation of parameters in large offspring number models and ratios of coalescence times. Theoretical Population Biology, 80: 16–28.

Eldon, B., Birkner, M., Blath, J., and Freund, F. 2015. Can the site-frequency spectrum distinguish exponential population growth from multiple-merger coalescents? Genetics, 199: 841–856.

Eldon, B., and Wakeley, J. 2006. Coalescent processes when the distribution of offspring number among individuals is highly skewed. Genetics, 172: 2621–2633.

Feller, W. 1971. An Introduction to Probability Theory and Its Applications, Volume 2, 2nd edn. John Wiley & Sons, New York. 669 pp.

Felsenstein, J. 1982. Numerical methods for inferring evolutionary trees. Quarterly Review of Biology, 57: 379–404.

Grant, W. S. 2015. Problems and cautions with sequence mismatch analysis and Bayesian skyline plots to infer historical demography. Journal of Heredity, 106: 333–346.

Griffiths, R. C. 2002. Ancestral inference from gene trees. *In* Modern Developments in Theoretical Population Genetics, pp. 94–117. Ed. by M. Slatkin, and M. Veuille. Oxford University Press, New York.

Guindon, S., Dufayard, J. F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Systematic Biology, 59: 307–321.

Hauser, L., and Carvalho, G. R. 2008. Paradigm shifts in marine fisheries genetics: ugly hypotheses slain by beautiful facts. Fish and Fisheries, 9: 333–362.

Hedgecock, D. 1994. Does variance in reproductive success limit effective population sizes of marine organisms? *In* Genetics and Evolution of Aquatic Organisms, pp. 122–134. Ed. by A. Beaumont. Chapman & Hall, London.

Hedgecock, D., and Pudovkin, A. I. 2011. Sweepstakes reproductive success in highly fecund marine fish and shellfish: a review and commentary. Bulletin of Marine Science, 87: 971–1002.

Hilborn, R., and Walters, C. J. 1992. Quantitative Fisheries Stock Assessment: Choice, Dynamics and Uncertainty. Chapman & Hall, London. 570 pp.

Hudson, R. R. 1983. Testing the constant-rate neutral allele model with protein sequence data. Evolution, 37: 203–217.

Hudson, R. R. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics, 18: 337–338.

Huillet, T. E. 2014. Pareto genealogies arising from a Poisson branching evolution model with selection. Journal of Mathematical Biology, 68: 727–761.

Ishikawa, S., Aoyama, J., Tsukamoto, K., and Nishida, M. 2001. Population structure of the Japanese eel *Anguilla japonica* as examined by mitochondrial DNA sequencing. Fisheries Science, 67: 246–253.

Janko, K., Lecointre, G., DeVries, A., Couloux, A., Cruaud, C., and Marshall, C. 2007. Did glacial advances during the Pleistocene influence differently the demographic histories of benthic and pelagic Antarctic shelf fishes?—inferences from intraspecific mitochondrial and nuclear DNA sequence diversity. BMC Evolutionary Biology, 7: 220.

Kingman, J. F. C. 1982. On the genealogy of large populations. Journal of Applied Probability, 19A: 27–43.

Kingman, J. F. C. 1993. Poisson Processes. Oxford University Press, New York. 104 pp.

Leaf, R. T., and Friedland, K. D. 2014. Autumn bloom phenology and magnitude influence haddock recruitment on Georges Bank. ICES Journal of Marine Science, 71: 2017–2025.

Niwa, H.-S. 2006a. Recruitment variability in exploited aquatic populations. Aquatic Living Resources, 19: 195–206.

Niwa, H.-S. 2006b. Exploitation dynamics of fish stocks. Ecological Informatics, 1: 87–99.

Niwa, H.-S. 2007. Random-walk dynamics of exploited fish populations. ICES Journal of Marine Science, 64: 496–502.

Nomura, S., Kobayashi, T., Agawa, Y., Margulies, D., Scholey, V., Sawada, Y., and Yagishita, N. 2014. Genetic population structure of the Pacific bluefin tuna *Thunnus orientalis* and the yellowfin tuna *Thunnus albacares* in the North Pacific Ocean. Fisheries Science, 80: 1193–1204.

Pinsky, M. L., and Palumbi, S. R. 2014. Meta-analysis reveals lower genetic diversity in overfished populations. Molecular Ecology, 23: 29–39.

Pitman, J. 1999. Coalescents with multiple collisions. Annals of Probability, 27: 1870–1902.

Riccioni, G., Landi, M., Ferrara, G., Milano, I., Cariani, A., Zane, L., Sella, M., *et al.* 2010. Spatio-temporal population structuring and genetic diversity retention in depleted Atlantic bluefin tuna of the Mediterranean Sea. Proceedings of the National Academy of Sciences of the United States of America, 107: 2102–2107.

Rodrigo, A. G., and Felsenstein, J. 1999. Coalescent approaches to HIV population genetics. *In* The Evolution of HIV, pp. 233–272. Ed. by K. A. Crandall. Johns Hopkins University Press, Baltimore.

Rogers, A. R., and Harpending, H. 1992. Population growth makes waves in the distribution of pairwise genetic differences. Molecular Biology and Evolution, 9: 552–569.

Sagitov, S. 1999. The general coalescent with asynchronous mergers of ancestral lines. Journal of Applied Probability, 36: 1116–1125.

Sargsyan, O., and Wakeley, J. 2008. A coalescent process with simultaneous multiple mergers for approximating the gene genealogies of many marine organisms. Theoretical Population Biology, 74: 105–114.

Sawyer, S. A., and Hartl, D. L. 1992. Population genetics of polymorphism and divergence. Genetics, 132: 1161–1176.

Schweinsberg, J. 2003. Coalescent processes obtained from supercritical Galton-Watson processes. Stochastic Processes and Their Applications, 106: 107–139.

Sjödin, P., Kaj, I., Krone, S., Lascoux, M., and Nordborg, M. 2005. On the meaning and existence of an effective population size. Genetics, 169: 1061–1070.

Slatkin, M., and Hudson, R. R. 1991. Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. Genetics, 129: 555–562.

Sromek, L., Lasota, R., and Wolowicz, M. 2015. Impact of glaciations on genetic diversity of pelagic mollusks: Antarctic *Limacina antarctica* and Arctic *Limacina helicina*. Marine Ecology Progress Series, 525: 143–152.

Steinrücken, M., Birkner, M., and Blath, J. 2013. Analysis of DNA sequence variation within marine species using Beta-coalescents. Theoretical Population Biology, 87: 15–24.

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

Tajima, F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics, 123: 585–595.

Wakeley, J., and Sargsyan, O. 2009. Extensions of the coalescent effective population size. Genetics, 181: 341–345.

Watterson, G. A. 1975. On the number of segregating sites in genetical models without recombination. Theoretical Population Biology, 7: 256–276.

Watterson, G. A., and Guess, H. A. 1977. Is the most frequent allele the oldest? Theoretical Population Biology, 11: 141–160.

Wright, S. 1931. Evolution in Mendelian populations. Genetics, 16: 96–159.

Zaliapin, I. V., Kagan, Y. Y., and Schoenberg, F. P. 2005. Approximating the distribution of Pareto sums. Pure and Applied Geophysics, 162: 1187–1228.

Handling editor: W. Stewart Grant