Population history of Antarctic and common minke whales inferred from individual whole-genome sequences

KISHIDA, Takushi
Wildlife Research Center, Kyoto University
2-24 Tanaka Sekiden-cho, Sakyo, Kyoto 606-8203, Japan
e-mail: taku.kishida@gmail.com

Keywords
Balaenoptera bonaerensis
Balaenoptera acutoarstrata
PSMC
Speciation in the open ocean has long been studied (e.g., Palumbi 1992, Miya and Nishida 1997, Davis et al. 2014, Friesen 2015), but it remains largely elusive how populations of highly mobile animals, such as whales, in such an open environment become reproductively isolated. Baleen whales of the genus *Balaenoptera* undertake extensive migrations, and there are few obvious barriers that potentially isolate their populations in the open ocean. Intra-species genetic diversities are extremely low in several species of these whales (Hoelzel 1994).

Among whales of the genus *Balaenoptera*, two species of minke whales, common minke whale *Balaenoptera acutorostrata* and Antarctic minke whale *B. bonaerensis*, are closely related (Pastene et al. 2007). Common minke whales are distributed worldwide, with subspecies identified in the North Atlantic (*B. acutorostrata acutorostrata*), North Pacific (*B. acutorostrata scammoni*), and Southern Ocean (dwarf minke whale). On the other hand, Antarctic minke whales are distributed only in the Southern Hemisphere. Minke whales are considered to have originated in the Southern Hemisphere, and common minke whales radiated into the North Atlantic and the North Pacific after speciation (Pastene et al. 2007). Based on the mitochondrial sequences, these two minke whale species are estimated to have diverged approximately 7.5 million years ago (MYA) (Park et al. 2015) to 4.7 MYA in the Southern Hemisphere (Pastene et al. 2007), and *B. a. acutorostrata* and *B. a. scammoni* diverged approximately 2.4 MYA (Park et al. 2015) to 1.5 MYA (Pastene et al. 2007). Pastene et al. (2007) argued that elevated ocean temperatures due to global warming in the Pliocene facilitated allopatric speciation into two minke whale species by disrupting the continuous belt of upwelling maintained by the Antarctic Circumpolar Current (ACC) in the Southern Hemisphere. They also speculated that an overall increase in carrying capacity as ACC-driven upwelling was re-established due to cooler temperature after the end of the Pliocene facilitated population expansion in both minke whale species.
The distribution of the time since the most recent common ancestor (TMRCA) between alleles in an individual gives information about the historical changes in effective population size ($N_e$) over time. Based on this, a revolutionary model for inferring demographic history from whole-genome sequence of a diploid individual, named pairwise sequential Markovian coalescent (PSMC), was proposed (Li and Durbin 2011). In this model, numerous independent loci contained in a diploid genome, each with its own TMRCA between the two alleles carried by an individual, are analyzed. When the PSMC inferences of two closely-related species that share a recent common ancestor are plotted together, the time point where the two historical effective population sizes diverge approximates the divergence time because these two species share the same $N_e$ value before divergence, though there are several caveats to be taken into account. For example, it is possible that the two populations had the same size after the split, which will lead to an underestimate (Prado-Martinez et al. 2013).

Recently, whole genome shotgun (WGS) reads of a North Pacific common minke whale sampled near the east coast of Korea were reported (Yim et al. 2014). In addition, my group sequenced WGS of an Antarctic minke whale from the Antarctic Ocean (Kishida et al. 2015). Using these data, historical $N_e$ of common and Antarctic minke whales were inferred with PSMC in order to further our understanding of demographic processes leading to the differentiation of these two species. Because all common minke whale subspecies belonged to the same ancestral population (i.e., same $N_e$) before divergence, the North Pacific sample of *B. a. scammoni* can be used to represent common minke whales for this comparison. Paired-end short reads of these whales listed in Table 1 were aligned to the BalAcu1.0 minke whale genome (Yim et al. 2014) with BWA (Li and Durbin 2009) after removal of adapter sequences and low quality bases using Trimomatic ver. 0.33 (Bolger et al. 2014) (Trimomatic settings: ILLUMINAACLIP:$HOME/Trimmomatic-0.33/adapters/TruSeq3-PE-2.fa:2:30:10 LEADING:20
Duplicate reads were removed and scaffolds identified as putative sex chromosome regions by Yim et al. (2014, see their Supplementary Table 16) were excluded from further analyses. Methods of Li and Durbin (2011) were followed to obtain the diploid consensus sequences, which excluded the following sites: (1) the read depth is more than twice or less than one-third of the average depth, (2) the site is within 5 bp around a predicted short insertion or deletion, and (3) the root-mean-square mapping quality is low (<10) (Table 1). Based on these consensus sequences, the PSMC outputs were obtained with 64 atomic time intervals and 28 free interval parameters (the ‘-p’ parameter used for PSMC was set to ‘4+25*2+4+6’). Bootstrapping was performed 100 times by breaking the consensus sequences into 5Mb segments and randomly resampling a set of segments by replacement.

The PSMC plots thus obtained are shown in Fig. 1. Average mutation rate per base per year (μ) and average generation time (g) are required for scaling results of PSMC, but it is generally not easy to estimate these values. These values may differ across species and might have been changed through evolution. Currently, the generation time g is estimated to be approximately 22 yr in both species of minke whales (Taylor et al. 2007). Regarding mutation rate, analyzing only limited numbers of intronic loci, it was previously approximated that μ=5×10^{-10} in baleen whales (Alter et al. 2007, Jackson et al. 2009). However recently, based on the genome-wide analyses, Yim et al. (2014) estimated that μ=1.07×10^{-9} in common minke whales. In addition, it is estimated that μ=1.22×10^{-9} in Yangtze River dolphins and μ=0.84×10^{-9} in common bottlenose dolphins (Zhou et al. 2013). Therefore, I assumed that μ is approximately 1×10^{-9} in whales. Based on these assumptions (g=22, μ=1×10^{-9}), historical N_e was estimated for both species (Fig. 1).
Yim et al. (2014) also estimated the demographic history of North Pacific common minke whales with PSMC and obtained essentially the same results. In addition to the historical Ne of common minke whales, I estimated that of Antarctic minke whales and plotted both together in Fig. 1.

Based on the PSMC results, I found that the effective population size of *B. a. scammoni* decreased profoundly about 1.5-3 MYA, when common minke whales are thought to have immigrated to the Northern Hemisphere and diverged into two subspecies (Pastene et al. 2007). Over the same period of time, the effective population size of Antarctic minke whales was also reduced. These findings may suggest that there had been environmental deterioration for minke whales around the Antarctic at that time. However, it is also possible that these inferred decrease of Ne are caused by population structuring (Mazet et al. 2016), especially by population divergence (Foote et al. 2016).

Unfortunately, probably due to lower heterozygosity in the North Pacific minke whale genome (Yim et al. 2014), Ne of common minke whales is not plotted where 2μT>6×10^-3 (before 3 MYA). This means that I could not find the exact point where *B. bonaerensis* and *B. acutorostrata* speciated, which is estimated to be approximately 4-5 MYA (Pastene et al. 2007). Prado-Martinez et al. (2013) proposed another approach to infer divergence time with PSMC. In this approach, they hybridized two haploid sequences from each species and then ran PSMC on the pseudo-diploid sequence thus obtained, and suggested that the time point where the inferred Ne goes to infinity corresponds to the divergence time. Following their methods, I generated the pseudo-diploid sequence of *B. bonaerensis* and *B. acutorostrata scammoni* using seqtk program provided by Heng Li (https://github.com/lh3/seqtk), and estimated the ancestral Ne with 5, 18, and 28 free interval parameters (Fig. 2). To derive haploid sequences from each minke whale species, an allele with low estimated consensus quality (<20) was excluded and
an allele at a heterozygous site was selected randomly (i.e., the ‘-rq20’ option was used for seqtk). As a result, there were sequences of 1.43 Gbp in total remained for PSMC calculation.

Fig. 2 shows that PSMC infers excessively large effective population size at around $2\mu T=2\times10^{-3}$-$3\times10^{-3}$ (1-1.5 MYA, when $\mu=1\times10^{-9}$) except in the case of 5 intervals. This is inconsistent with the previous report that *B. bonaerensis* and *B. acutorostrata* diverged some 4-5 MYA (Pastene *et al.* 2007), even if the mutation rate is smaller than my assumption (e.g., 2-3 MYA, when $\mu=5\times10^{-10}$, as suggested by Alter *et al.* 2007 and Jackson *et al.* 2009). Following these results, there are two hypotheses to be taken into account. First, two species of minke whales diverged at approximately $2\mu T=3\times10^{-3}$ (1.5 MYA), much more recently than the previous estimations. Indeed, PSMC plots of Antarctic and common minke whale shown in Fig. 1 seem to diverge at around $2\mu T=3\times10^{-3}$. And second, it is because there had been gene flow between these two minke whale species after speciation. Pastene *et al.* (2007) speculated that both *B. bonaerensis* and *B. acutorostrata* were distributed in the Southern Hemisphere before the divergence of *B. a. acutorostrata* and *B. a. scammoni* at around 1.5 MYA, and this inconsistency in the estimated divergence time of two minke whale species may be caused if there had been inter-species gene flow between populations of *B. bonaerensis* and *B. acutorostrata* when they lived sympatrically in the Southern Hemisphere (Fig. 3A). Following the second hypothesis, a model of population divergence mimicking the evolution of Antarctic and common minke whales with no changes in the total size of $N_e$ is considered (Fig. 3B).

Based on this model, PSMC plots of populations 1 and 2, generated using ms-HOT program (Hellenthal and Stephens 2007) modified by Heng Li (msHOT-lite, https://github.com/lh3/foreign/tree/master/msHOT-lite ), are shown in Figs 3C (gene flow is not assumed) and 3D (symmetric gene flow is assumed). In Fig. 3D, sizes of both populations 1 and 2 are inferred to be decreased gradually at around $2\mu T = 2\times10^{-3}$ to $5\times10^{-3}$ (1-2.5 MYA). The
same tendency is also observed in Fig. 1. These data suggest that the second hypothesis is also possible, that is there had been gene flow between populations of Antarctic and common minke whales when they lived in the Southern Hemisphere together (4.7-1.5 MYA), and that the apparent decrease in the inferred effective population sizes at this point in evolution was mainly caused by population divergence. In any case, after that, $N_e$ of Antarctic minke whales increased, while there was not a profound increase in the effective population size of North Pacific minke whales. Such population history should contribute to current census population sizes of these whales —there are approximately 500,000-700,000 Antarctic minke whales in the Southern Hemisphere, whereas there are less than 30,000 common minke whales in the North Pacific (Thomas et al. 2016). Further analyses, including North Atlantic common minke whales and dwarf minke whales, will reveal speciation and divergence of minke whale populations more clearly.

Acknowledgements

I am grateful to Dr. Kentaro Fukuta (National Institute of Genetics) and reviewers of this manuscript for helpful comments. Computations were partially performed on the NIG supercomputer at ROIS National Institute of Genetics. This study was financially supported in part by JSPS KAKENHI (grant nos. 24770075 and 15K07184).
Figure 1

Demographic history of minke whales inferred with pairwise sequential Markovian coalescent. Bootstrap replicates are represented by light narrow lines. The x axis gives time, and the upper x axis indicates scaling in years with the assumption of $\mu=1\times10^{-9}$. The y axis gives the effective population size, and scaling in numbers with the assumption of $\mu=1\times10^{-9}$ and $g=22$ is shown in right.

Abbreviations and parameters: MYA; million years ago, KYA; kilo years ago, $T$; time before present [year], $N_e$; effective population size, $\mu$; mutation rate per base per year, $g$; generation time [year].

Figure 2

Pairwise sequential Markovian coalescent (PSMC) plots of a pseudo-diploid genome sequence of *B. bonaerensis* and *B. acutorostrata scammoni*. The ‘-p’ parameter used for PSMC is set to ‘20+4*5’, ‘6+17*2’ and ‘4+25*2+4+6’ for 5, 18, and 28 intervals, respectively.

Figure 3
A) A phylogenetic tree of minke whales taken from Pastene et al. (2007). Minke whale subspecies which were not analyzed in this study are represented by light narrow lines. Because both *B. bonaerensis* and *B. acutorostrata* distributed in the Southern Hemisphere sympatrically approximately 4.7 to 1.5 MYA (Pastene *et al.* 2007), it is possible that there had been gene flow between these two whale populations at that time.

B) A model of population divergence mimicking the evolution of Antarctic and common minke whales with no changes in the size of populations, which in total remains 40,000. An ancestral population of size $N_{A1}=40,000$ splits 214,000 generations ago into two descending populations of size $N_1=N_{A2}=20,000$, and the population of size $N_{A2}=20,000$ splits 68,200 generations ago into two descending populations of size $N_2=N_3=10,000$. Populations 1, 2, and 3 mimic the populations of *B. bonerensis*, *B. a. scammoni*, and *B. a. acutorostrata*, respectively (dwarf minke whale population is not considered).

Based on this model, with the assumption of $\mu=1\times10^{-9}$ and $g=22$, demographic history of populations 1 and 2 were inferred with pairwise sequential Markovian coalescent (C, D). Gene flow between populations was not assumed in (C), whereas it was assumed that there is a period of symmetric gene flow (migration rate per generation $m=0.001$) after divergence until the split of populations 2 and 3 (214,000-68,200 generations ago) in (D). Each data set consists of 100 blocks of 10 Mb (total of 1 Gb) sampled from a single diploid individual, generated with the following command lines: (C) population 1, msHOT-lite 2 100 -l -t 8800 -r 7040 10000000 -l 3 2 0 0.0 -n 1 2 -n 2 1 -n 3 1 -G 0.0 -ej 1.705 3 2 -en 1.705 1 2 -en 1.705 2 2 -ej 5.35 2 1 -en 5.35 1 4; population 2, msHOT-lite 2 100 -l -t 8800 -r 7040 10000000 -l 3 2 0 0.0 -n 1 1 -n 2 2 -n 3 1 -G 0.0 -ej 1.705 3 1 -en 1.705 1 2 -en 1.705 2 2 -ej 5.35 2 1 -en 5.35 1 4, and (D) population 1, msHOT-lite 2 100 -l -t 8800 -r 7040 10000000 -l 3 2 0 0.0 -n 1 2 -n 2 1 -n 3 1 -G 0.0 -m 1 2 0 -m 2 -em 1.705 1 2 40 -em 1.705 2 1 40 -en 1.705 1 2 -en 1.705 2 2 -ej 5.35 2 1 -em


Effective population size (scaled in units of $4\mu gN_e \times 10^3$)

Time (scaled in units of $2\mu T$)

- B. bonaerensis
- B. acutorostrata scammoni
A) B. bonaerensis (Antarctic)
B. acutorostrata (Antarctic)
B. a. acutorostrata (North Atlantic)
B. a. scammoni (North Pacific)

B) 214,000 generations ago (4.7 MYA)
68,200 generations ago (1.5 MYA)

C) Effective population size (scaled in units of $4\pi g N_e \times 10^3$)
Time (scaled in units of $2\mu T$)

D) Effective population size (scaled in units of $4\pi g N_e \times 10^3$)
Time (scaled in units of $2\mu T$)
Table 1. Sequence data used for PSMC.

<table>
<thead>
<tr>
<th>species</th>
<th>Sequence Read Archive accession nos.</th>
<th>average depth(^1)</th>
<th>callable(^2)</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>B. acutorostrata scammoni</em></td>
<td>SRR893003, SRR896642, SRR901891</td>
<td>51.1</td>
<td>1.49Gbp</td>
</tr>
<tr>
<td><em>B. bonaerensis</em></td>
<td>DRR014695</td>
<td>82.2</td>
<td>1.45Gbp</td>
</tr>
</tbody>
</table>

\(^1\) Average depth was calculated after removal of duplicated reads and sex chromosomal reads

\(^2\) Total length of regions for calculation. Following sites are excluded: (1) the read depth is more than twice or less than one-third of the average depth, (2) the site is within 5 bp around a predicted short indel, and (3) the root-mean-square mapping quality is low (<10).