SC/F19/WNPM/02

Collaborative analysis of WNP minke whale stock structure using the Japanese microsatellite DNA database and spatially explicit population structure analyses

Menno de Jong and A. Rus Hoelzel

INTERNATIONAL
WHALING COMMISSION

Collaborative analysis of WNP minke whale stock structure using the Japanese microsatellite DNA database and spatially explicit population structure analyses.

Menno de Jong & A. Rus Hoelzel

Department of Biosciences, Durham University, South Road, Durham, DH1 3LE, UK

Abstract:

Our objective was to help address the recommended 'analysis 2' from the report of the workshop on Western North Pacific common minke whale stock structure (SC_67B_REP-05) in support of the next inter-sessional meeting on WNP common minke whale stock structure using the microsatellite DNA database provided by the Institute of Cetacean Research. This specific aspect of the work applied spatially explicit population structure analyses that provide greater power than the program STRUCTURE. The data were analysed as a total dataset (not based on any assignment in STRUCTURE), and included temporal subdivision to assess possible seasonal changes in patterns of connectivity. We compared a subset of samples genotyped at 26 loci with the same samples at 16 loci and determined that the larger sample size provided for the 16-locus dataset provided greater power than the increased number of loci for the smaller sample set, and so further analyses were done based on 16 loci. We ran Geneland, TESS and BAPS and found the former most informative. Additional analyses were conducted to test the inference from Geneland that suggested four putative populations in the dataset.

Introduction:

Stock structure for the common minke whale in the Western North Pacific has been investigated in IWC committees for a number of years in preparation for the implementation review, and intensively through the NPM working group of the RMP subcommittee. At SC62 there were a dozen papers including reviews by Pastene et al. (SC62/NMP12) and Wade & Baker (SC62/NPM15). All recognised a clear distinction between a 'J' stock found primarily in the Sea of Japan and an 'O' stock in the Pacific. Additional potential stocks identified at the time included a 'W' (also referred to as O_E) stock identified sporadically in sub-area 9 based on mtDNA haplotypes, a 'Y' stock in the Yellow Sea and possibly along the eastern coast of Korea, and a coastal stock along the eastern coast of Japan, possibly distinct in the southern area (J_E) and northern area (O_W). Support for an O_W stock came from 'haplogroup' frequency data (after collapsing haplotype frequency into 4 groups based on concern about the very large number of singleton sequences; Baker et al. SC62/NPM20). At the same time, comparisons using mtDNA based on F_{ST} and Chi-square looking for substructure in either O-stock or J-stock regions found no significant results (Park et al. SC62/NPM21). Further discussions were held during an inter-sessional workshop in Korea and at SC63.

From SC63 a table and text was generated by Hoelzel, Waples & Gaggiotti (Appendices 8&9 in SC63 Annex D1) that summarised the data to date and considered the various components of evidence (including genetic data) for putative stocks. Evidence was strong for J and O stocks and is not further reviewed here. The case for Y was deemed moderate based on significant but small microsatellite DNA F_{ST} values (comparing SA5 and the Sea of Japan), and on seasonal evidence for mixing based on HWE deviation (Kanda et al. SC62/NPM11). The case for J_E was considered low, because although there was a small significant mtDNA F_{ST} comparing areas 6E and 2, it was not clear if this reflected mixing of J

and O in area 2. The case for O_W was considered moderate, based on single-locus Wahlund effects suggesting more than a simple mixture of J and O (though more mixture models were needed for greater certainty; Waples SC63/RMP7), an individual-based PCA analysis that involved various levels of sample purging (Gaggiotti & Gascuel SC63/RMP23), and the earlier work based on haplotype frequencies (Baker et al. SC62/NPM20). The plausibility of a multi-stock scenario was promoted again at SC64 by Wade & Baker (SC/64/NPM11).

In general, there were many permutations of comparisons based on assignment in STRUCTURE (Pritchard et al. 2000), designating 'pure' O and J together with 'unassigned' samples, not fully reviewed here. However, STRUCTURE has relatively low power, and so it is hard to determine from that analysis alone whether unassigned individuals reflect admixture between J and O or one or more separate stocks. When plotting the location of unassigned individuals Pastene et al. (SC/F16/JR38) found that many clustered in the coastal regions of area 7, though sporadically in other areas as well, and the proportion was similar across all areas. Most recently, Pastene et al. (SC/F16/JR/40) used 16 microstaellite DNA loci and applied heterogeneity tests and DAPC to investigate possible substructure in the O-stock. They found no significant patterns and concluded that there was no substructure, however this analysis was again based on a purged (O-only) dataset, subsequent to analysis using STRUCTURE. Tiedemann et al. (RS6747_SC_SDDNA_01) reported parent offspring matches based on a false discovery rate of 0.1 and 16 microsatellite loci, and found matches between area 7C with areas 2, 8 and 9, suggesting gene flow between these sub-areas of the O stock. At the intersessional workshop in Japan (see SC_67B_REP-05), it was concluded that additional analyses should be undertaken to further address the question of possible substructure in J and O stocks, and in particular, the nature of the individuals unassigned in STRUCTURE. This was to include an assessment of the power added when 26 rather than 16 loci were included, temporal and spatial analyses that included all samples, the removal of 'J-stock' individuals using DAPC (followed by the analysis of the rest purged of J-stock whales), and the spatially explicit analyses undertaken here. As agreed at the workshop, we used three available methods for spatial genetic analyses: Geneland (Guillot et al. 2005), TESS (see Caye et al. 2016) and BAPS (Corander et al. 2008), with an emphasis on Geneland.

Materials & Methods:

Input data

Our input dataset consisted of capture locations (longitude and latitude) and microsatellite DNA genotype scores for 4707 minke whales from the western North Pacific. These samples were collected between 1994 and 2016 from management sub-areas 1E, 2C, 6E, 7CN, 7CS, 7WR, 8, 9, 10E and 11. We removed all individuals that had a parent present in the dataset (as determined by earlier parental analyses and identified by our collaborators at ICR), retaining 4656 individuals, of which 519 individuals were genotyped for 26 loci and the remaining 4139 individuals for only 16 of the 26 loci. We ran initial comparisons of a 519-sample dataset based on 16 compared to 26 loci, and all other analyses were based on the full dataset of 4656 individuals or on temporal or spatial subsets.

Geneland

Geneland is a landscape genetics program run in R that clusters samples into homogeneous groups assuming approximate Hardy-Weinberg and linkage equilibrium (HWLE), and incorporating spatial data. We ran Geneland_4.0.8 trials with 0.1M, 1M, 3M, 5M or 10M MCMC iterations. Convergence of the MCMC chain was assessed using visual examination of the traceplot as well as with the geweke diagnostics implemented in the R

package 'coda'. Final runs were executed using 1M iterations and a burn-in of 0.2M iterations, as these settings resulted in good convergence of the MCMC chain, not clearly improved by longer runs. We executed Geneland on the Durham Hamilton mainframe cluster by running set-up commands in R (version 3.4.4; see Appendix 1). An initial run set 'varnp' to 'TRUE', meaning that the analysis would estimate the best fit for the number of populations (K). Additional runs started with a fixed K. The allele frequency model was set to either 'correlated' or 'uncorrelated'. A parameter that considers the accuracy of location data ('delta-coord') was set to 0 (accurate locations). The MCMC analysis was thinned to every 100 records for display and analysis. The initial run was for correlated allele frequencies where varnp=TRUE. Subsequent runs using the full dataset were either uncorrelated, correlated with a fixed K, or repeats. Most runs were based on 16 loci, but runs based on 8 loci were also included.

TESS

TESS is a Bayesian clustering program using tessellations (division of samples into best fit polygons) to define populations. We ran Tess2.3 on a Windows7 computer using the CAR admixture model, for K varying between 2 and 5, with 3 runs per K, 100000 sweeps and a burn-in of 20000 sweeps, and with the 'update spatial interaction'-parameter enabled.

BAPS

BAPS uses Bayesian models to capture genetic population structure by describing the molecular variation in each subpopulation using a separate joint probability distribution over the observed loci. We ran Baps6.0 on a Windows7 computer using two modules: 'Spatial clustering of individuals' and 'Spatial clustering of groups'. We ran BAPS in the 'non-fixed K' mode with a maximum K of 5. We also ran BAPS with fixed K.

Analysis of clusters defined by Geneland

Given that the strongest pattern of substructure found was based on analyses in Geneland, we investigated only these putative populations further using summary statistics, ordination and Bayesian analyses. Summary statistics F_{ST} and $\Delta\mu^2$ were generated in Arlequin 3.5 (Excoffier & Lischer, 2010) comparing population divisions identified in Geneland, as well as among geographic populations that reflected those divisions. A subset of samples (first 100 in the list from each putative population) were compared by factorial correspondence analysis (FCA) run in Genetix 4.04 (Belkhir et al. 2002). The reduced sample size was analysed to facilitate visual interpretation, and the program was run using the 'sur la population' option. Since this analysis assigns individuals to a position in Euclidian space based on their multilocus genotypes, the subsample should provide a reasonable representation of the full dataset. Directional gene flow was estimated using Bayesian inference implemented in BayesAss 3.0-SNPs (see Wilson & Rannala 2003). Parameter settings in BayesAss were 1 million iterations, a burn-in of 100,000 generations, a seed of 10 and a delta value of 0.1. Allele frequency distributions were compared at each locus using Spearman's rank and Peterson's correlations. These computations and illustrations were run in R (see Appendix 1). A preliminary assessment using approximate Bayesian analyses (ABC) using the package DIYABC (Cornuet et al. 2008) was included based on sub-samples from putative populations (N=200 per sub-population) using flat priors and assuming no admixture.

Plotting output data

In addition to the standard output plots created by Geneland, TESS and BAPS, we generated illustrations using in-house R scripts incorporating geographical maps with sample

locations and cluster assignment indicated by colour coding. Our R scripts work from input matrices (qmatrices) of Bayesian posterior cluster assignment probabilities. The qmatrices are standard output from each of the programs used in this study. Further detail is provided in Appendix 1.

Results

Geneland

Initial trial runs compared the resolution of K=2 (assuming correlated allele frequencies) using 519 samples when 16 or 26 loci were used to generate the analysis in Geneland (Fig. 1). The resolution was better when 26 loci were used, but the populations could be clearly separated with 16 loci, and the resolution with 16 loci increased greatly when the number of samples was increased from 519 to 4656 (Fig. 2). We therefore based all further analyses on 16 loci using the full dataset of 4656 whales (after removal of calves from known mother-offspring pairs), together with some comparisons using 8 loci (allowing a further assessment of the power gained from multiple loci). We then addressed the question of whether the allele frequencies should be considered correlated or uncorrelated. When run as uncorrelated, the best supported K=3 (Fig. 3). Convergence for this run appears stable (Fig. 3a) and the three putative populations map strongly to accepted distributions for J-stock and O-stock, but also indicate a coastal stock in sub-areas 7CN, 7CS and 11 (see Fig, 3d). We then tested the pattern of correlation of allele frequencies between samples assigning clearly to the J-stock and O-stock region (Fig. 4). Although allele frequency profiles diverge at some loci, overall the correlation is strong (Pearson's = 0.891; Spearman's = 0.805). We therefore focus on analyses based on the assumption of correlated allele frequencies (implying some level of admixture).

When the full sample-set was included under the correlated allele frequency model the best supported K was 4 (Fig. 5). As before, there appeared to be good MCMC convergence (Fig. 5a), and apart from samples assigned to the geographic regions expected for J-stock and O-stock there was additional assignment to samples from sub-areas 7CN, 7CS and 11, this time divided into two sub-groups (Fig. 5d). Repeat runs showed essentially the same clustering, even when only 8 loci were used (data not shown). Figure 6 facilitates the visual assessment of the distributions by separating out the individual populations by colour. When individual months were run separately and assessed K=2, but given the differential distribution of samples in different months and the smaller sample sizes, it was not possible to determine which pair of putative populations was being depicted. Instead we assigned individuals to groups based on the full data set, and then considered the distribution of those identified groups when plotted month by month (Fig. 7). This suggested that one putative coastal population in sub-areas 7 and 11 (depicted in red in Fig.s 5-7) was in relatively consistent geographic positions throughout most of the year. The other (blue in Fig.s 5-7) appeared to move northward through the course of the year (Fig. 7). When we ran a fixed K=4, two of these clusters overlapped such that the best estimate was actually K=3, including putative populations consistent with J-stock, O-stock and a coastal population mostly in the 7CS, 7CN, & 11 sub-areas (Figure 8).

We then compared the four putative populations identified by Geneland (using the varnP function to identify the number of clusters) to consider their potential nature and origin. We used a measure of genetic distance designed for microsatellite DNA data ($\Delta\mu^2$) as a descriptive measure to quantify the relative amount of divergence between the groups identified in Geneland (Table 1). We show these same comparisons using F_{ST} because this is a familiar metric, and not to assess population structure, as that has already been determined by Geneland. We then compared the relevant geographic sub-areas by $\Delta\mu^2$ and F_{ST}, and

while samples within 7C are likely mixed, they remain significantly differentiated from sub-areas representing J-stock and O-stock (Table 2), though the level of differentiation is small. Factorial correspondence analysis was run from a subset of samples from each putative group (the first 100 samples from each list), to facilitate visual discrimination among clusters (Fig. 9). While there is extensive overlap, there is some differentiation between all four clusters, depending on the combination of factors compared. Factor 1 explained 72.6%, factor 2 15.0% and factor 3 12.5% of the variance.

Contemporary directional gene flow among these four putative populations was estimated using BayesAss (Fig. 9, Tables 4&5). Gene flow is relatively low between the 'orange' (O-stock) and 'green' (J-stock) populations and between the two putative coastal populations in this analysis. Estimated gene flow between orange and green is highly directional into the two coastal populations (see illustration in Fig. 10a). When the sub-areas represented by the putative populations identified in Geneland are compared, the inference is similar, but strong directional migration from one of the putative coastal populations ('red') into O-stocks is now suggested, which could be due to the presence of O-stock animals in the coastal areas (Fig. 9b).

The preliminary analysis using ABC compared four evolution scenarios (Fig. 11a), where the putative populations are coded by colour as in figures 5-7. None of the prior distributions for these scenarios overlapped with the observed data (Fig. 11b), suggesting that greater complexity will be required (especially admixture) for the comparator set of scenarios. Among the four scenarios tested, scenario 1 was best supported according to the logistic regression, though none received strong support from the direct regression (Table 5). From scenario 1, the estimated effective size of the populations was largest in putative J and O-stock regions, as expected (Table 6). The 95% confidence interval for division times between either stock J or O and the putative coastal stocks did not overlap with zero, and the mean values ranged from ~1000 to ~20,000 ybp, compared to an estimated division time of ~900,000 ybp between J and O stocks (Table 6).

TESS

Two repeat runs using the same parameters gave slightly differing results, but each detected 3 clusters, one representing mostly 'J-stock' whales, one mostly O-stock, and a third in the coastal regions of sub-areas 7CN, 7CS and 11 (see Fig. 12). The assignment overlap with Geneland for the coastal population varied from 73-77% overlap, though only ~8% of those assigned to coastal populations in Geneland were also assigned to the coastal population in TESS.

BAPS

Both module runs found K=2 as the most likely outcome, and the distribution of samples was a moderate fit to the J and O stock distributions, with fairly extensive overlap (see Fig. 13). We also ran BAPS in the 'fixed K-mode' for K = 3, which resulted in the program assigning all but one sample to two clusters (results not shown).

Discussion

Spatially explicit analyses run in Geneland and TESS consistently identified putative stocks J and O as expected, but also found one or more putative populations primarily in areas 7CS, 7CN and 11 (with some also found in area 2C). BAPS identified only 2 putative populations, but these were mixed in areas 7, 8 and 9, suggesting relatively poor assignment. We did not pursue this method beyond the initial runs. For TESS and Geneland the number of putative coastal populations found depended on the program and on the parameter settings.

TESS only identified one putative coastal population, and while there was overlap with coastal populations identified by Geneland (73-77% of those identified in TESS also identified in Geneland), TESS only identified ~8% of the individuals identified to the coastal populations by Geneland. When Geneland was run assuming uncorrelated allele frequencies, or when set to a fixed value for K, three populations were found (including a coastal population found primarily in areas 7CS, 7CN and 11). A separate analysis comparing allele frequency distributions between likely J and O-stock whales suggested strong correlation, and so we focussed on analyses that assumed correlated allele frequencies, and let Geneland determine the best supported value for K in that context. These runs split the putative coastal population identified by the previous runs in Geneland into two sub- populations, supporting a total of 4 populations.

Subsequent analyses were designed to test possible inference about the nature and origin of these putative coastal populations. They appear in the same sub-areas (mostly 7CS, 7CN and 11) that had been suggested to contain possible additional stocks based on analyses using PCA (Gaggiotti & Gascuel SC63/RMP23), Wahlund effects (Waples SC63/RMP7) and mtDNA haplogroups (Baker et al. SC62/NPM20). Summary statistics suggested considerably less differentiation between putative coastal populations and those identified as representing J and O-stocks than between J and O-stock sample sets (Table 1). When the approximate geographic areas represented by the Geneland 'colours' were compared, all F_{ST} values were significant, but comparisons between coastal (7CS & 7CN or 10E and 11) and O-stock (7WR, 7E, 8 & 9) were much smaller than between putative O-stock and J-stock (2C, 1E & 6E) areas. When analysed by FCA using a subset of samples, there was good separation between the clusters representing putative J and O-stock whales, and fairly extensive overlap with the putative costal population clusters. However, there were individuals within each coastal cluster that did not overlap with J and O clusters (depending on the combination of factors compared), and these coastal clusters overlapped fairly evenly with both J and O clusters (Fig. 9). This may suggest admixture with both, rather than each reflecting miss-assigned individuals from J and O stocks, respectively. The possibility of admixture with both J and O stocks in each putative coastal population was supported by the analysis in BayesAss (Fig. 10), where strong directional gene flow was indicated from both J and O stocks into each coastal population (rather than e.g. just gene flow between J and the 'red' coastal population or O and the 'blue' coastal population).

Since some analyses suggested K=3, it is unclear whether the two putative coastal populations are legitimately distinct. The genetic distance between them was small, but not the smallest value (Table 1). From the distribution of samples over time based on temporal comparisons there is some indication of distinct patterns of movement between the two putative coastal populations, with the 'blue' population appearing to move northward during the course of the year, but this is hard to quantify and may be sample dependent (Fig. 7). From the preliminary ABC analysis the best supported model had the coastal populations evolving from separate source populations (see Table 5 & Fig. 11), but this analysis will require further work to improve the representation of putative scenarios (see Fig. 10b). Another key question is whether or not the putative coastal stocks are real, or simply a representation of the poorly assigned (potentially admixed) individuals at the interface between the parapatric J and O stocks. From the ABC analysis, the 95% confidence estimated division times between the coastal populations and either J or O-stock did not overlap with zero (Table 6), but results from this analysis should be interpreted cautiously until a more robust analysis can be undertaken. This should be based on the inclusion of more complex scenarios (based on admixture), and the full dataset, but the computational time required will be fairly extensive. The difference between the pattern of gene flow indicated from comparing the putative Geneland populations, in contrast to the comparison

among regional populations suggests that the latter reflects a mixing of coastal and O-stock whales in the coastal areas, inconsistent with them being from the same stock. This pattern was not seen with J-stock, again suggesting the presence of an additional stock in the 7C regions.

There may be precedent for a coastal resident stock of minke whales in the eastern North Pacific where long-term photo-identification studies found fidelity of whales to local habitat in local waters, with some individuals observed throughout most of the year and for multiple years (e.g. Dorsey et al. 1990). Given that there have been multiple indications of possible stock structure within sub-areas 7CS, 7CN and 11 in the past (see above), these additional indications from this study suggest that a precautionary approach would be appropriate involving a distinct management protocol for this region. However, the data suggest whales in coastal populations from this region would be either very recently diverged from or fairly extensively admixed with (or both) the more fully isolated J and O-stock populations. One possible scenario would be the evolution of coastal resident populations derived from each of the J and O-stock parent populations in association with environmental change during the Holocene.

Acknowledgements

We thank Luis Pastene, Mioko Taguchi and the IRC for access to their extensive databased of microsatellite DNA genotypes. Many thanks to Fatih Sarigoel and Andrea Estandia for their assistance with setting up runs on the HPC cluster.

References

Baker CS, Hammer RM, Brownell RL, Wade PR (2010) Stock structure of western North Pacific minke whales based on mtDNA haplogroups from 'bycatch' and scientific whaling. SC62/NPM20

Belkhir, K, Borsa, P, Chikhi, L, Raufaste, N, Bonhomme, F (2002). GENETIX 4.04, logiciel sous Windows TM pour la g_en_etique des populations. Montpellier, France: Laboratoire G_enome, Populations, Interactions.

Caye K, Deist TM, Martins H, Michel O, François O (2016) TESS3: Fast inference of spatial population structure and genome scans for selection. Molecular Ecology Resources 16, 540-548.

Corander J, Marttinen P, Sirén J, Tang J. (2008) Enhanced Bayesian modelling in BAPS software for learning genetic structures of populations. BMC Bioinformatics, 9, 539.

Cornuet JM, Santos F, Beaumont MA, Robert CP, Marin JM, Balding DJ, Guillemaud T, Estoup A (2008) Inferring population history with DIYABC: a user-friendly approach to Approximate Bayesian Computations. Bioinformatics, 24, 2713-2719.

Dorsey EM, Stern SJ, Hoelzel AR, Jacobsen J (1990) Minke whales (*Balaenoptera acutorostrata)* from the west coast of North America: individual recognition and small-scale site fidelity (SC/A88/ID21)

Excoffier, L, Lischer, HEL (2010) Arlequin suite v. 3.5: A new series of programs to perform population genetics analyses under Linux and Windows. Molecular Ecology Resources, 10, 564–567.

Gaggiotti O, Gascuel F (2011) Stock structure of North Pacific minke whales as revealed by principal component analyses of microsatellite data. SC63/RMP23

Guillot G, Mortier, F, Estoup, A (2005) Geneland: A program for landscape genetics. Molecular Ecology Notes, 5, 712-715.

Kanda N, Park JY, Goto M, An Y-R, Choi S-G, Moon D-Y, Kishiro T, Yoshida H, Kato H, Pastene LA (2010) Genetic analysis of western North Pacific minke whales from Korea and Japan based on microsatellite DNA. SC62/NMP11

Park JY, Goto M, Kanda N, Kishiro T, Yoshida H, Kato H, Pastene LA (2010) Mitochondrial DNA analyses of J and O stocks common minke whale in the western North Pacific. SC62/NPM21

Pastene LA, Goto M, Kanda N (2010) Progress in the development of stock structure hypotheses for western North Pacific common minke whales. SC62/NPM12

Pastene LA, Goto M, Taguchi M, Kitakado T (2018) Temporal and spatial distribution of the J and O stocks of common minke whale in waters around Japan based on microsatellite DNA. SC/F16/JR/38

Pastene LA, Goto M, Taguchi M, Kitakado T (2018) Updated genetic analyses based on mitochondrial and microsatellite DNA indicated no sub-structure of the O-stock common minke whale in the western North Pacific. SC/F16/JR/40

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

Tiedemann R, Tiedemann MR, Goto M, Taguchi M, Pastene LA (2017) Finding parent-offspring pairs among western North Pacific common minke whales. SC/67A/SDDNA/01

Wade PR, Baker CS (2010) A review of the plausible range of stock structure hypotheses of western North Pacific minke whales using genetic and other biological information. SC62/NPM15.

Wade PR, Baker CS (2012) Hiplausibility of hypothesis III for stock structure of the western North Pacific minke. SC64/NPM11

Waples RS (2011) Can evidence for spatial and/ or temporal genetic heterogeneity of North Pacific minke whales be explained by different kixture fractions of the same two core stocks, or is it necessary to postulate an additional stock (s)? SC63/RMP7

Wilson, GA, Rannala, B (2003). Bayesian inference of recent migration rates using multilocus genotypes. Genetics, 163, 1177–1191.

Table 1: Upper diagonal is $\Delta\mu^2$, and lower diagonal is F_{ST}, all assigned according to Geneland output using the correlated allele frequency model for which K=4. All F_{ST} values are significant at the $p<0.001$ level.

	Red	Green	Blue	Orange
Red		0.0353	0.0231	0.0753
Green	0.0054		0.1003	0.1989
Blue	0.0068	0.0237		0.0209
Orange	0.0173	0.0407	0.0025	

Table 2: Upper diagonal is $\Delta\mu^2$, and lower diagonal is F_{ST}, all assigned according to management areas. All F_{ST} values are significant at the $p<0.001$ level.

	OW (7CS & 7CN)	**J** (2C, 1E & 6E)	**O** (7WR, 7E, 8 & 9)	**JN** (10E & 11)
OW		0.52272	0.08493	0.24466
J	0.0202		1.00726	0.26508
O	0.00326	0.03952		0.48793
JN	0.00315	0.00693	0.01237	

Table 3: BayesAss migration table, where 0=red, 1=green, 2=blue and 3=orange populations from the Geneland correlated allele frequency run. Migration is from the population on the right to the one on the left, and the standard error of estimates is given parenthetically.

m[0][0]:	0.6690(0.0022)	m[0][1]:	0.2090(0.0073)	m[0][2]:	0.0122(0.0065)	m[0][3]:	0.1098(0.0066)
m[1][0]:	0.0010(0.0009)	m[1][1]:	0.9922(0.0017)	m[1][2]:	0.0037(0.0017)	m[1][3]:	0.0031(0.0017)
m[2][0]:	0.0013(0.0012)	m[2][1]:	0.1036(0.0063)	m[2][2]:	0.7264(0.0666)	m[2][3]:	0.1687(0.0687)
m[3][0]:	0.0016(0.0016)	m[3][1]:	0.0110(0.0021)	m[3][2]:	0.0536(0.0282)	m[3][3]:	0.9338(0.0286)

Table 4: BayesAss migration table, where 0=east coast (areas 7CS & 7CN), 1=west coast (areas 2C, 1E & 6E), 2=Hokkaido (areas 10E and 11) and 3=Pacific Ocean (areas 7WR, 7E, 8 & 9) populations from the Geneland correlated allele frequency run. Migration is from the population on the right to the one on the left, and the standard error of estimates is given parenthetically.

m[0][0]:	0.8088(0.0225)	m[0][1]:	0.1113(0.0138)	m[0][2]:	0.0001(0.0001)	m[0][3]:	0.0798(0.0101)
m[1][0]:	0.0092(0.0038)	m[1][1]:	0.9816(0.0023)	m[1][2]:	0.0002(0.0002)	m[1][3]:	0.0090(0.0033)
m[2][0]:	0.1335(0.0242)	m[2][1]:	0.1681(0.0147)	m[2][2]:	0.6689(0.0022)	m[2][3]:	0.0295(0.0193)
m[3][0]:	0.2634(0.0894)	m[3][1]:	0.0020(0.0013)	m[3][2]:	0.0004(0.0004)	m[3][3]:	0.7343(0.0893)

Table 5: Output from the ABC analysis showing regression support values for scenarios associated with the direct and logistic methods.

Model	closest	scenario 1	scenario 2	scenario 3	scenario 4
Direct	50	0.3000 [0.0000,0.7017]	0.3000 [0.0000,0.7017]	0.1800 [0.0000,0.5168]	0.2200 [0.0000,0.5831]
	100	0.2800 [0.0000,0.6736]	0.2300 [0.0000,0.5989]	0.2000 [0.0000,0.5506]	0.2900 [0.0000,0.6877]
	150	0.2200 [0.0000,0.5831]	0.2667 [0.0000,0.6543]	0.2400 [0.0000,0.6144]	0.2733 [0.0000,0.6640]
	200	0.2300 [0.0000,0.5989]	0.2650 [0.0000,0.6518]	0.2550 [0.0000,0.6370]	0.2500 [0.0000,0.6296]
	250	0.2360 [0.0000,0.6082]	0.2360 [0.0000,0.6082]	0.2840 [0.0000,0.6793]	0.2440 [0.0000,0.6205]
	300	0.2500 [0.0000,0.6296]	0.2400 [0.0000,0.6144]	0.2767 [0.0000,0.6688]	0.2333 [0.0000,0.6041]
	350	0.2571 [0.0000,0.6402]	0.2314 [0.0000,0.6011]	0.2800 [0.0000,0.6736]	0.2314 [0.0000,0.6011]
	400	0.2625 [0.0000,0.6482]	0.2250 [0.0000,0.5910]	0.2825 [0.0000,0.6771]	0.2300 [0.0000,0.5989]
	450	0.2644 [0.0000,0.6510]	0.2333 [0.0000,0.6041]	0.2667 [0.0000,0.6543]	0.2356 [0.0000,0.6075]
	500	**0.2600 [0.0000,0.6445]**	**0.2340 [0.0000,0.6051]**	**0.2740 [0.0000,0.6649]**	**0.2320 [0.0000,0.6020]**
Logistic	**n=5005**	**0.9204 [0.8825,0.9584]**	**0.0010 [0.0000,0.0037]**	**0.0756 [0.0390,0.1121]**	**0.0030 [0.0009,0.0051]**

Table 6: Output from the ABC analyses showing parameter estimations for the best supported scenario (scenario 1 – see Figure 11a). Ne estimates colour-coded as in Figure 11.

	mean	median	mode	q025	q050	q250	q750	q950	q975
N1	7.78E+03	8.03E+03	8.36E+03	4.26E+03	4.92E+03	6.96E+03	8.94E+03	9.75E+03	9.87E+03
N2	9.80E+03	6.15E+03	4.15E+03	1.78E+03	2.17E+03	4.00E+03	1.03E+04	3.06E+04	4.20E+04
N3	3.33E+04	2.89E+04	2.19E+04	1.11E+04	1.29E+04	2.13E+04	4.09E+04	6.95E+04	7.87E+04
N4	9.28E+03	9.42E+03	9.57E+03	7.95E+03	8.22E+03	9.02E+03	9.68E+03	9.94E+03	9.97E+03
t1	1.12E+03	7.32E+02	3.09E+02	**1.10E+02**	1.50E+02	3.96E+02	1.25E+03	3.06E+03	**4.41E+03**
t2	1.88E+04	1.08E+04	3.52E+03	**1.36E+03**	1.70E+03	5.28E+03	2.34E+04	6.81E+04	**7.99E+04**
t0	9.13E+05	9.49E+05	9.82E+05	**5.91E+05**	7.08E+05	8.92E+05	9.79E+05	9.96E+05	**9.98E+05**

Figure 1: Output from Geneland for K=2 comparing 519 samples using 26 or 16 loci.

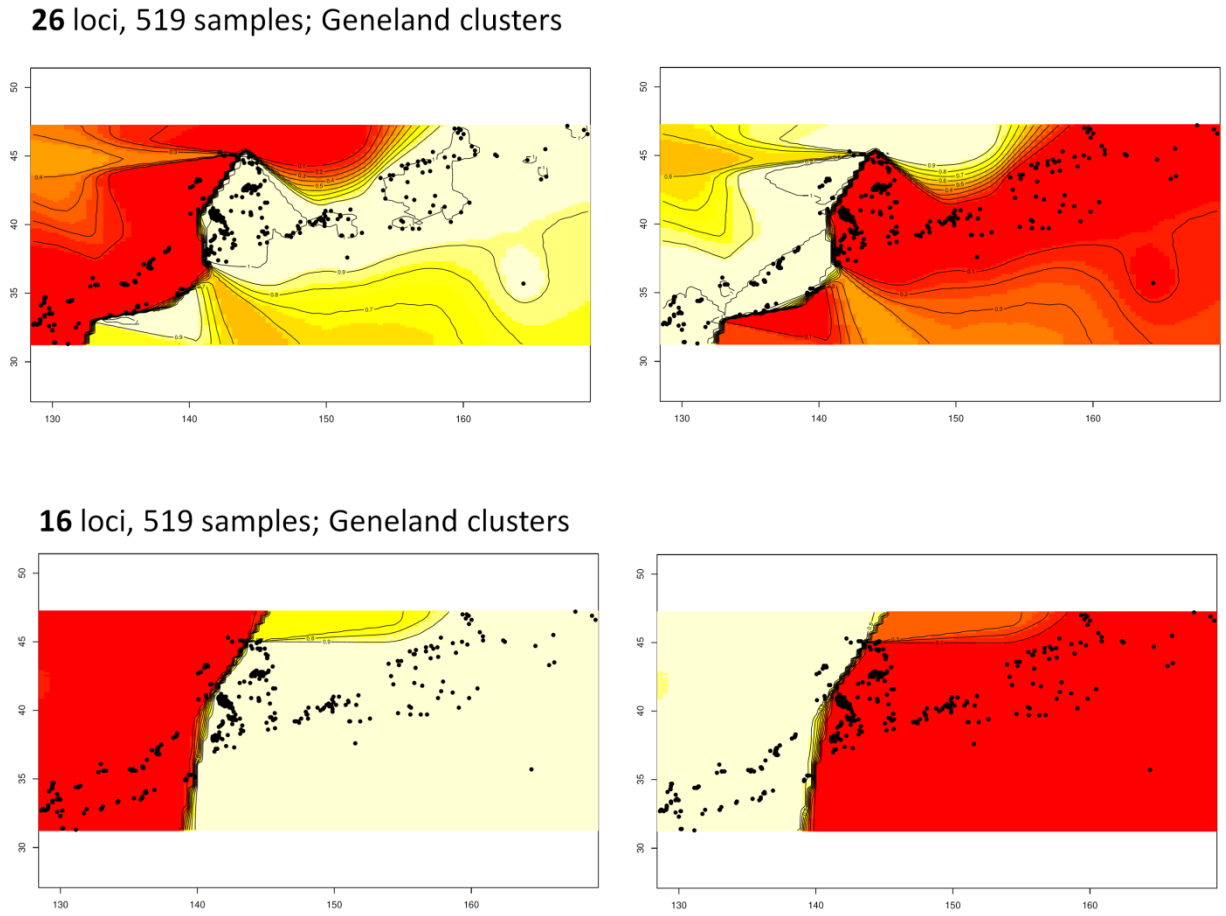


Figure 2: Output from Geneland for K=2 comparing 519 vs 4656 samples at 16 loci.

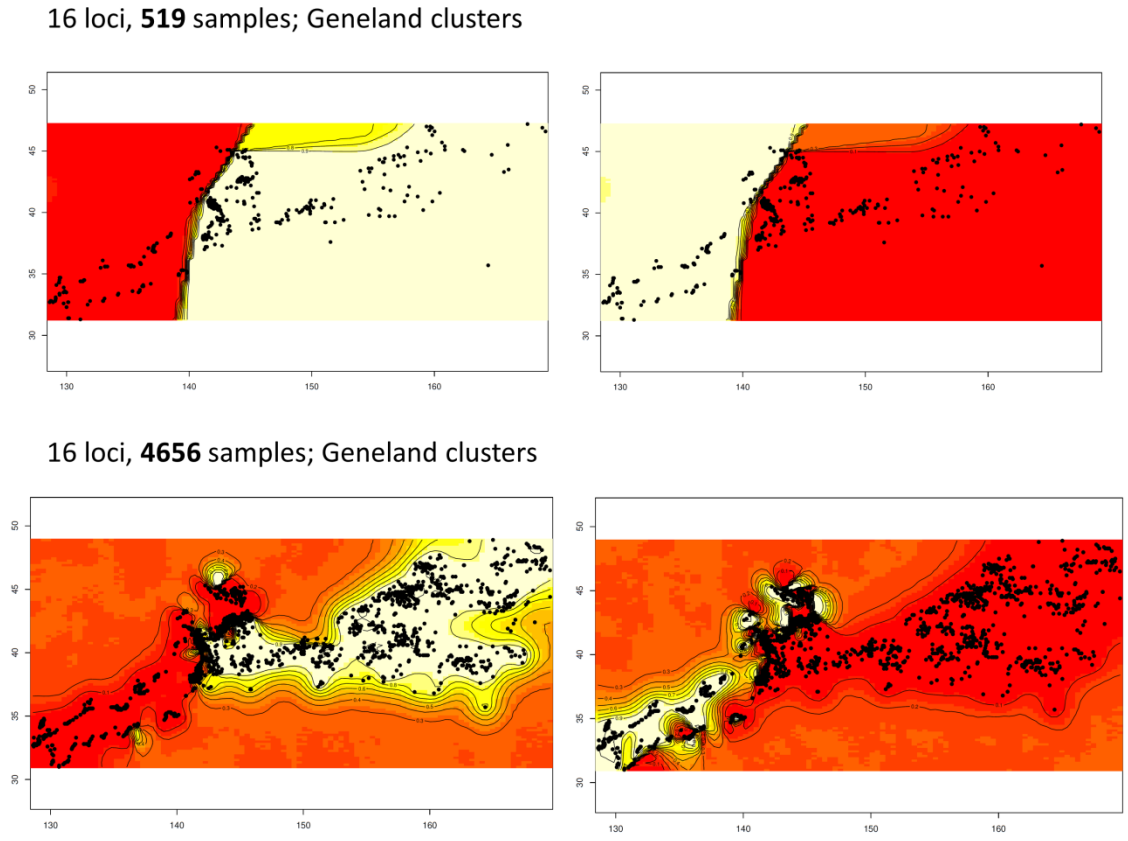


Figure 3: Geneland output for the uncorrelated allele frequency mode. a) The MCMC trace record, recorded with thining every 100 runs. b) The bandwidth and distribution of posterior outcomes. c) Posterior support for different values of K. d) The geographic distribution of samples assigned to the three clusters.

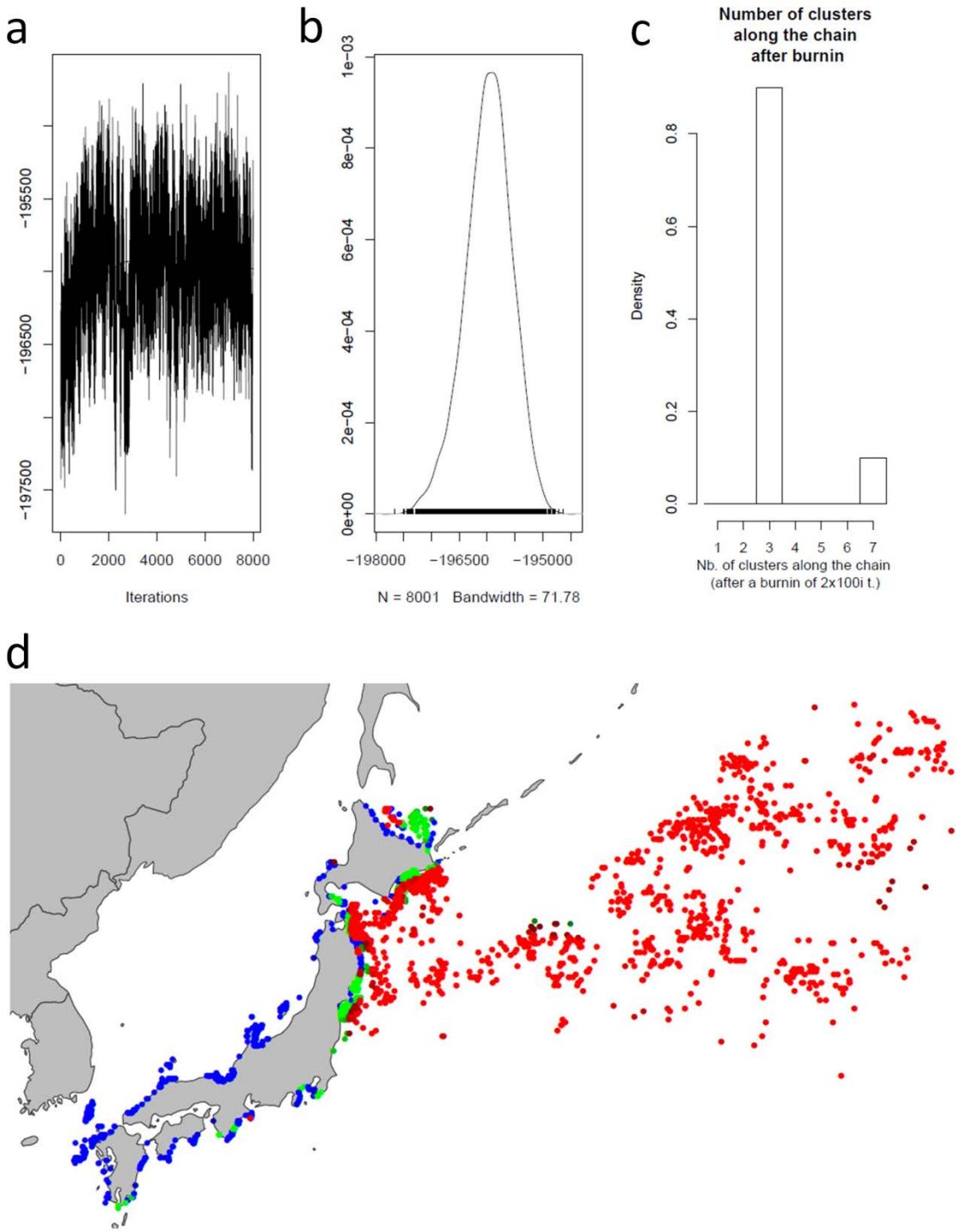


Figure 4: Allele frequency comparison between putative J-stock (blue) and O-stock (green) populations at 16 loci.

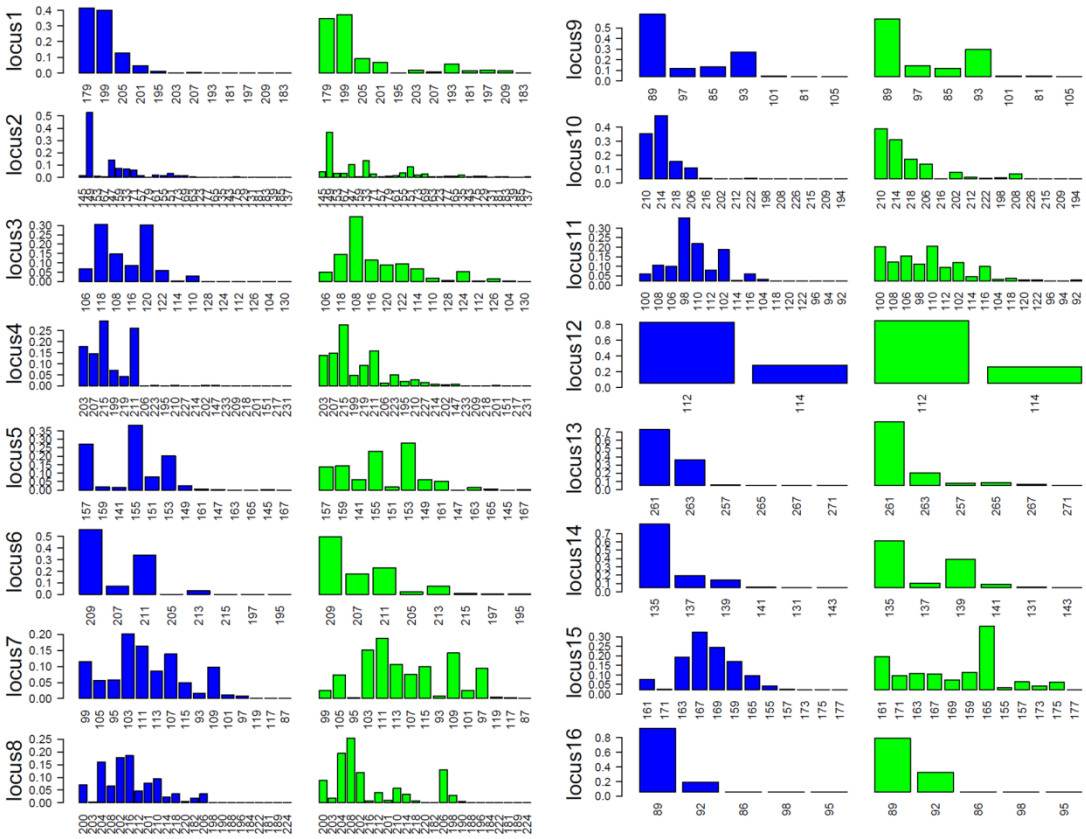


Figure 5: Geneland output for the correlated allele frequency mode. a) The MCMC trace record, recorded with thining every 100 runs. b) The bandwidth and distribution of posterior outcomes. c) Posterior support for different values of K. d) The geographic distribution of samples assigned to the four clusters.

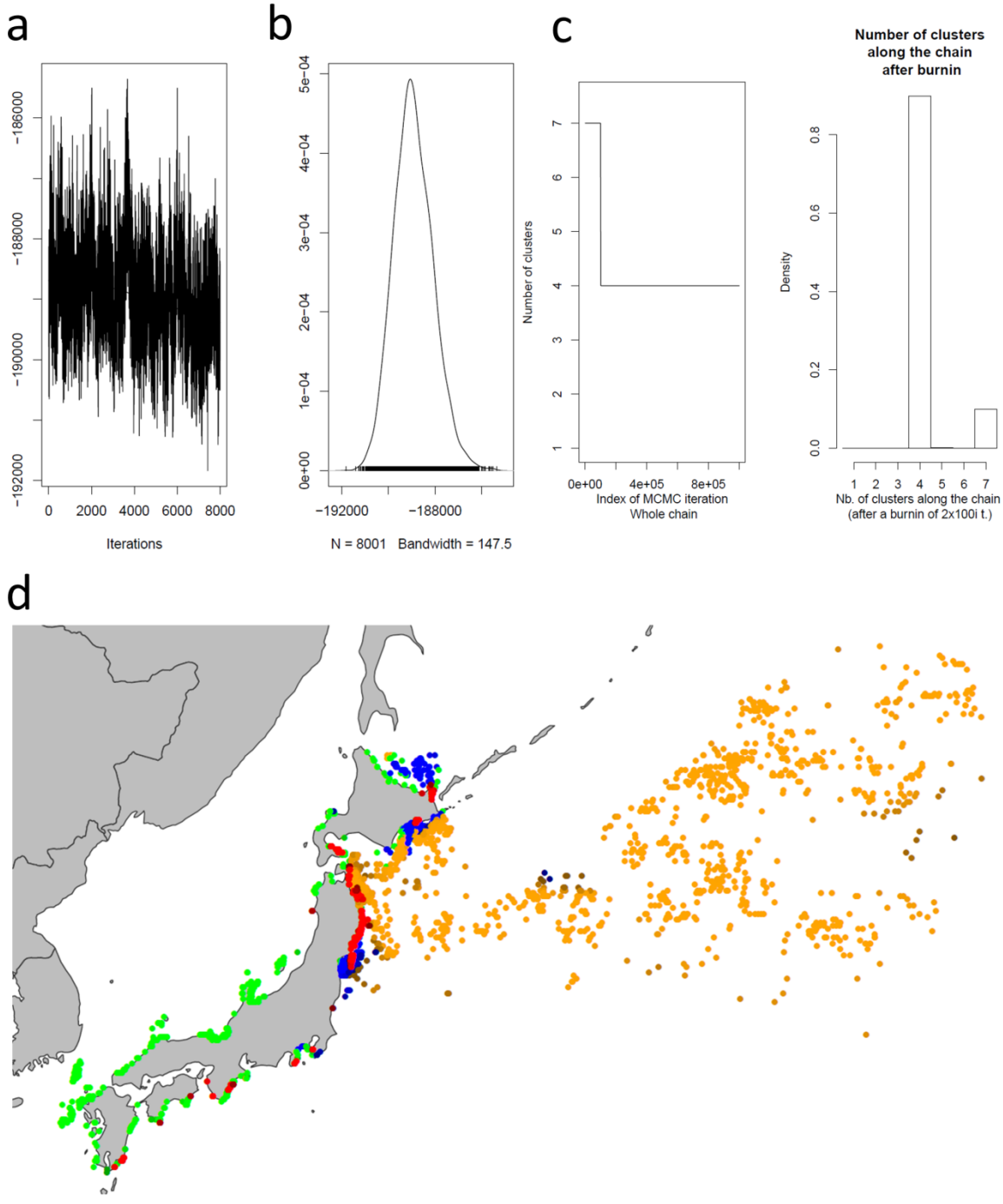


Figure 6: illustration of the putative populations one at a time.

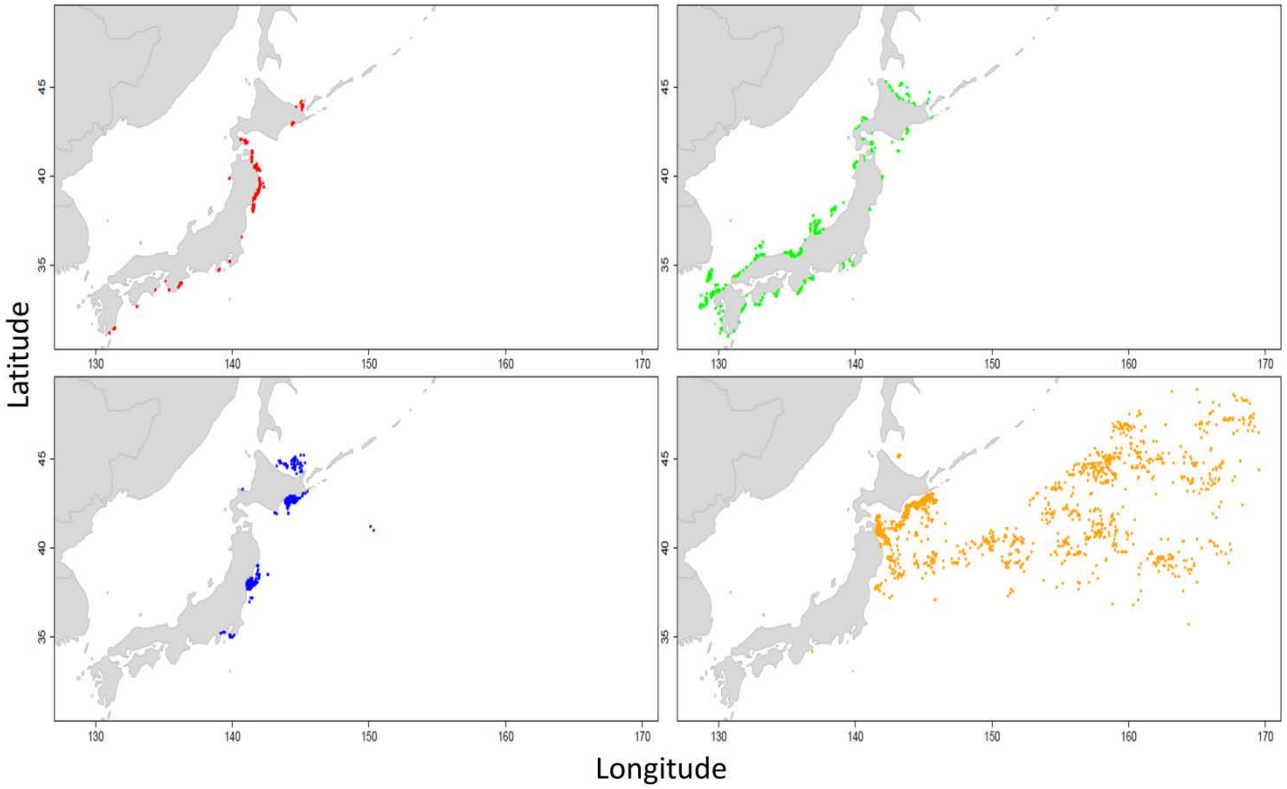


Figure 7: Geneland output by month based on the correlated allele frequency model.

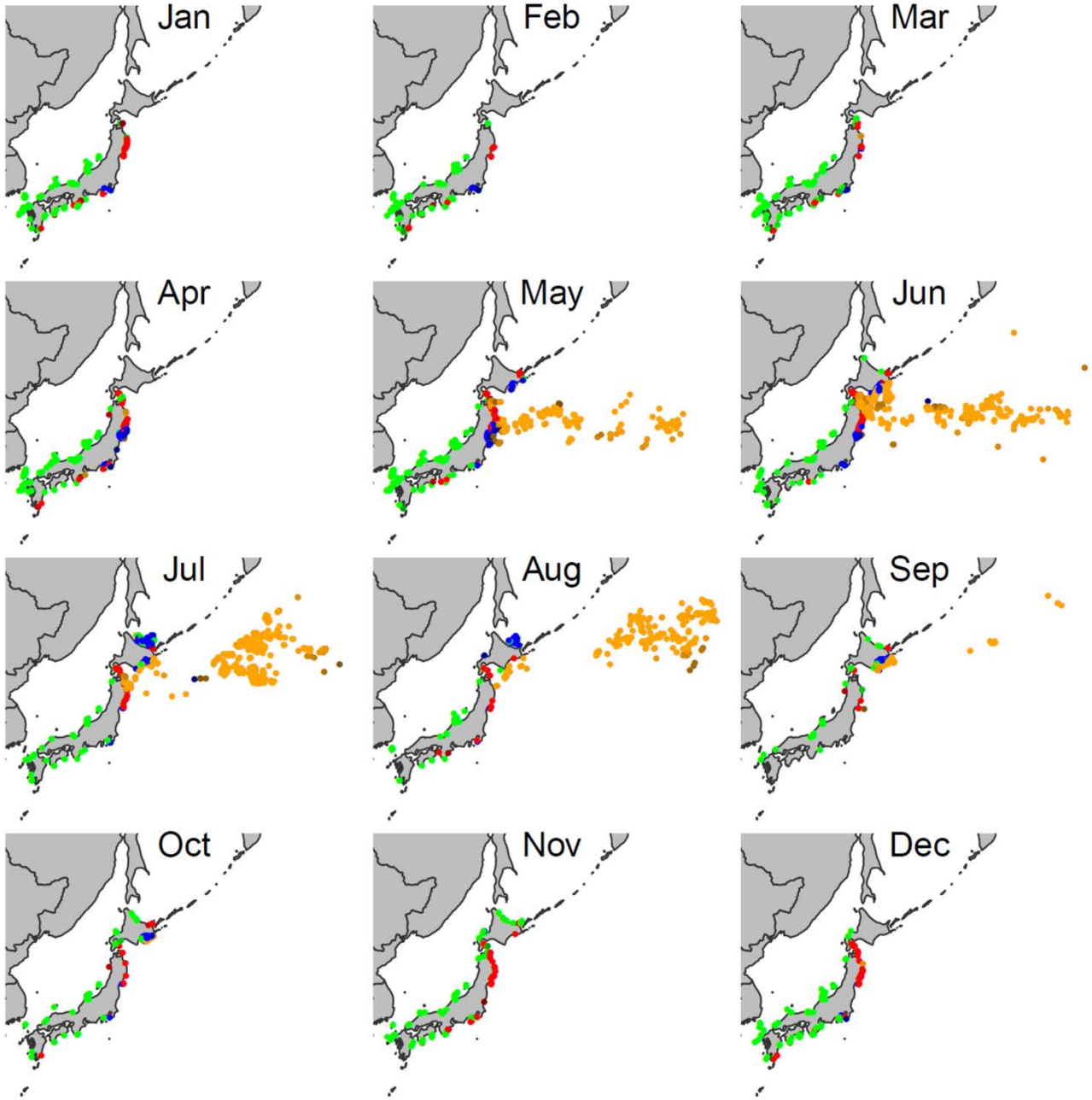


Figure 8: Geneland output when the correlated allele frequency model is applied with a fixed K=4.

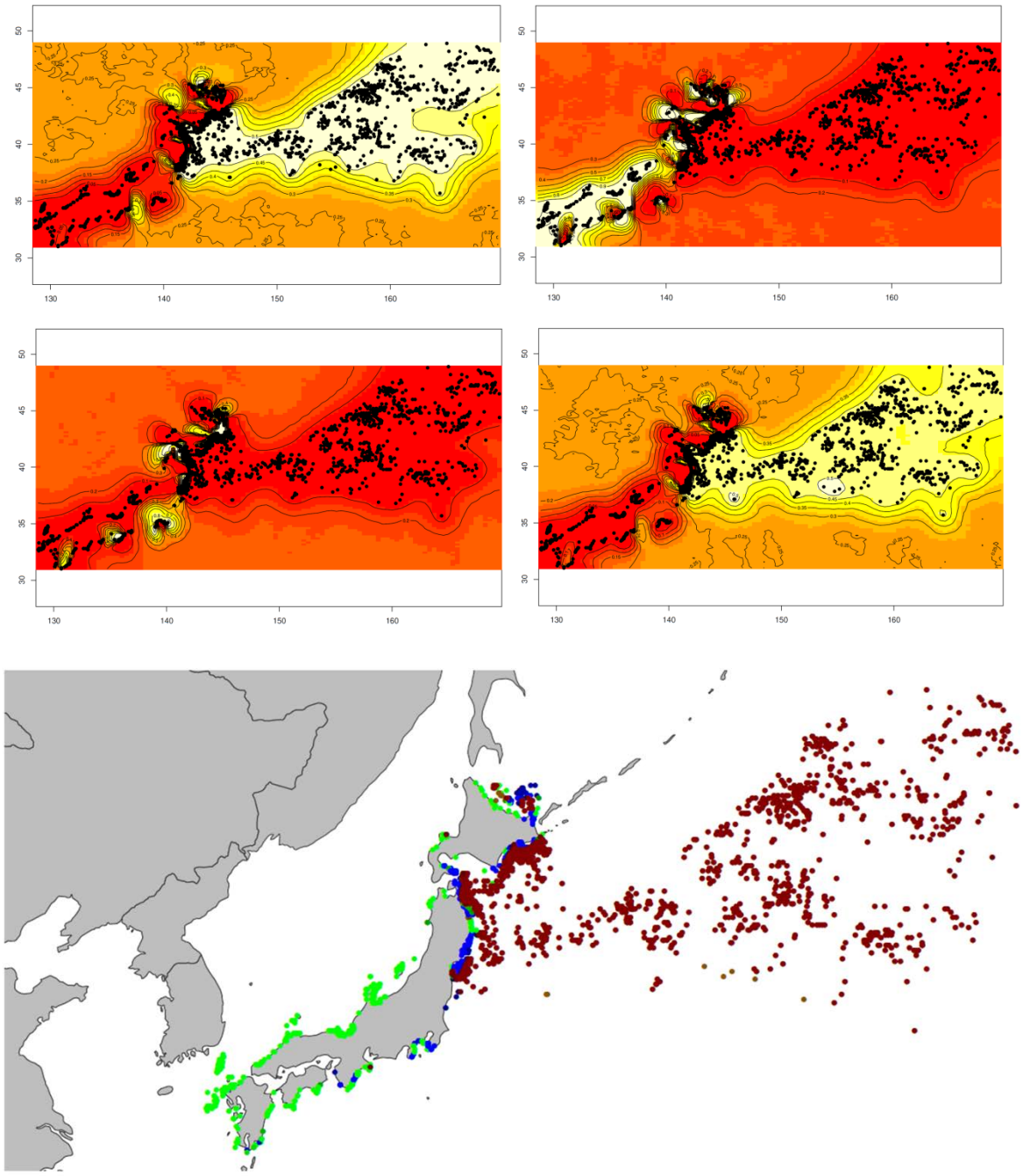
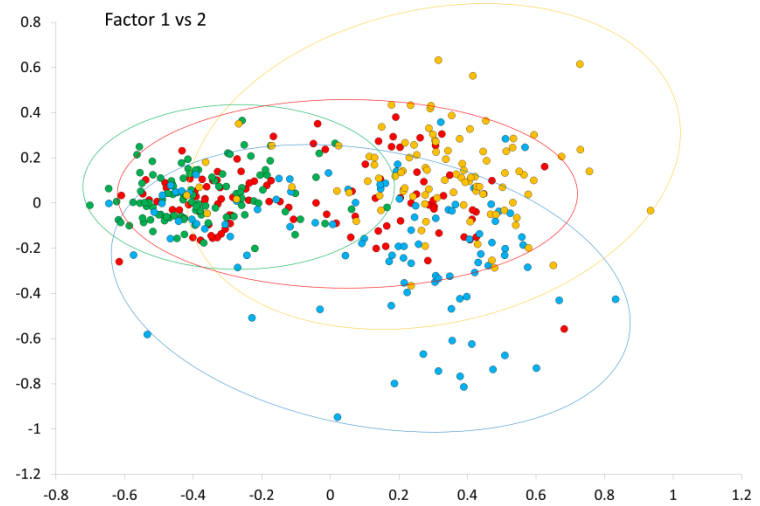
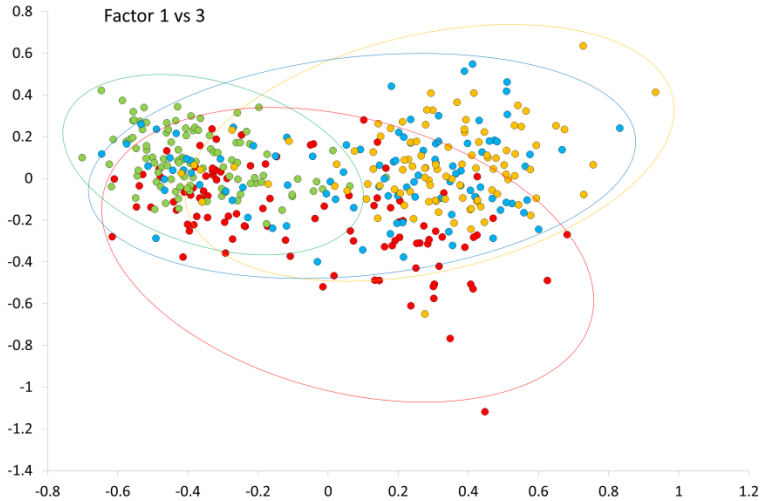


Figure 9: FCA comparison among putative populations as defined by Geneland based on the correlated allele frequency model (colour coding as in Figures 5-7). a) factors 1 vs 2, b) factors 1 vs 3; c) factors 2 vs 3.

a)



b)



c)

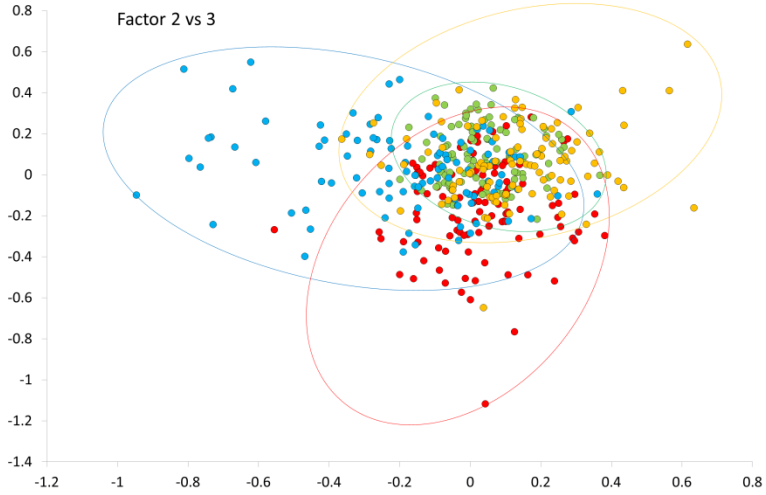
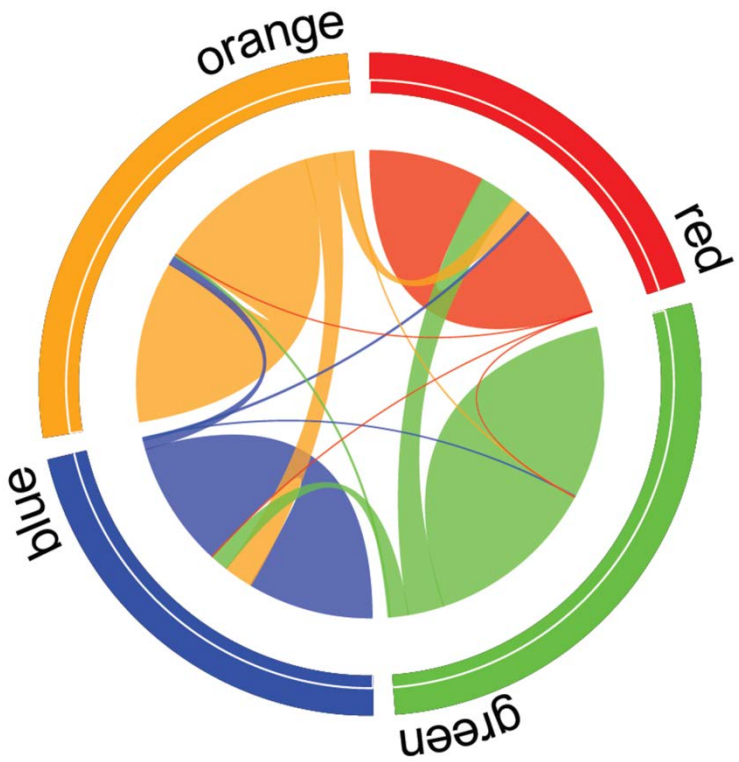


Figure 10: BayesAss assessment of contemporary gene flow assigning: a) populations based on the assignments found in Geneland (correlated allele frequency model), and b) populations based on location assignments, where ‘ocean’ is areas 7WR, 7E, 8 & 9, ‘east coast’ is 7CS & 7CN, ‘westcoast’ is 2C, 1E & 6E, and ‘Hokkaido’ is 10E and 11.

a)



b)

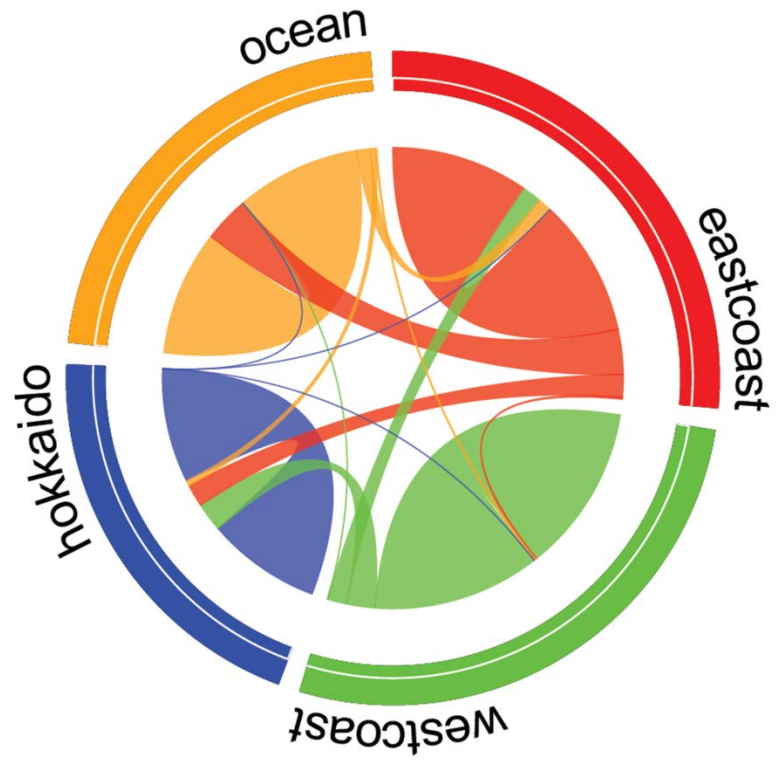


Figure 11: a) Scenarios assessed in the ABC analyses, using the colour-coding as illustrated in figures 5-7. b) A PCA assessment of the overlap between prior distributions for the four scenarios in comparison with the observed data.

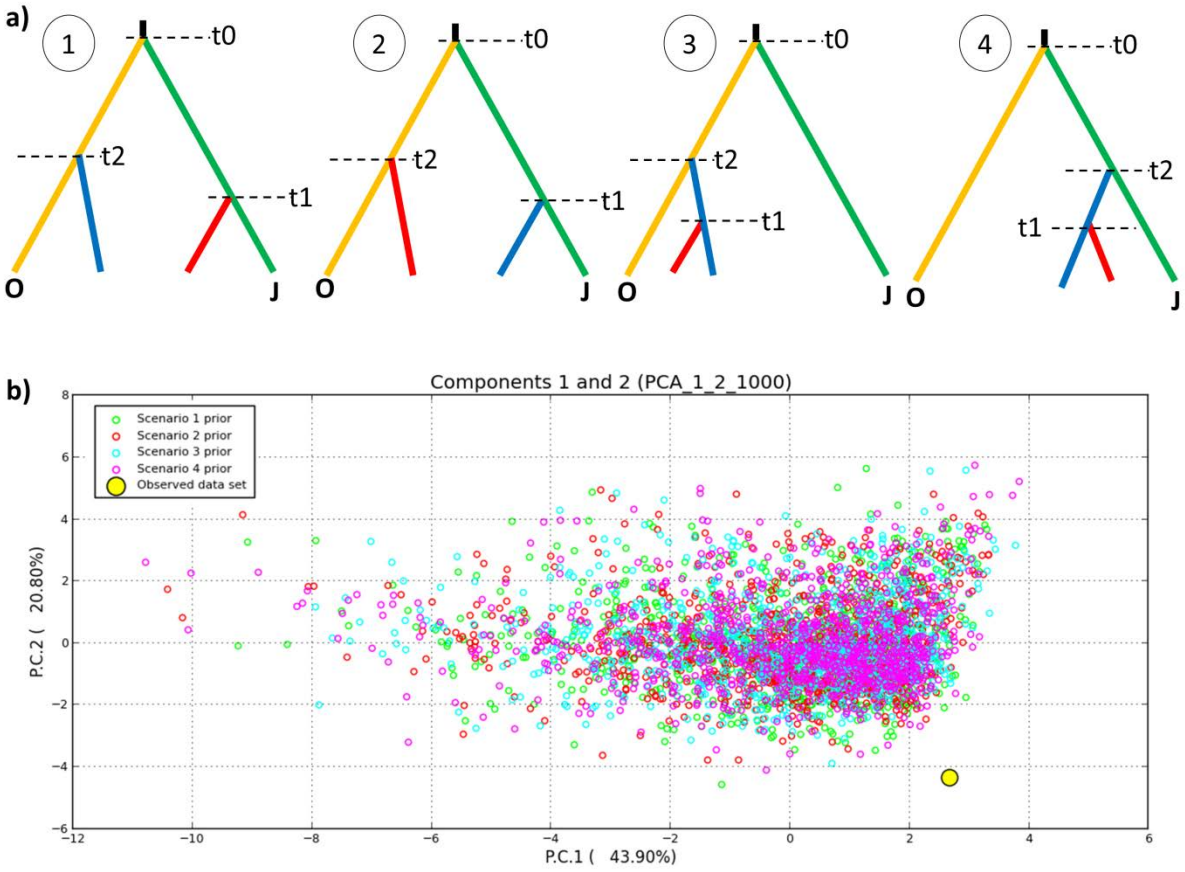


Figure 12: Two runs in TESS showing population assignment by colour.

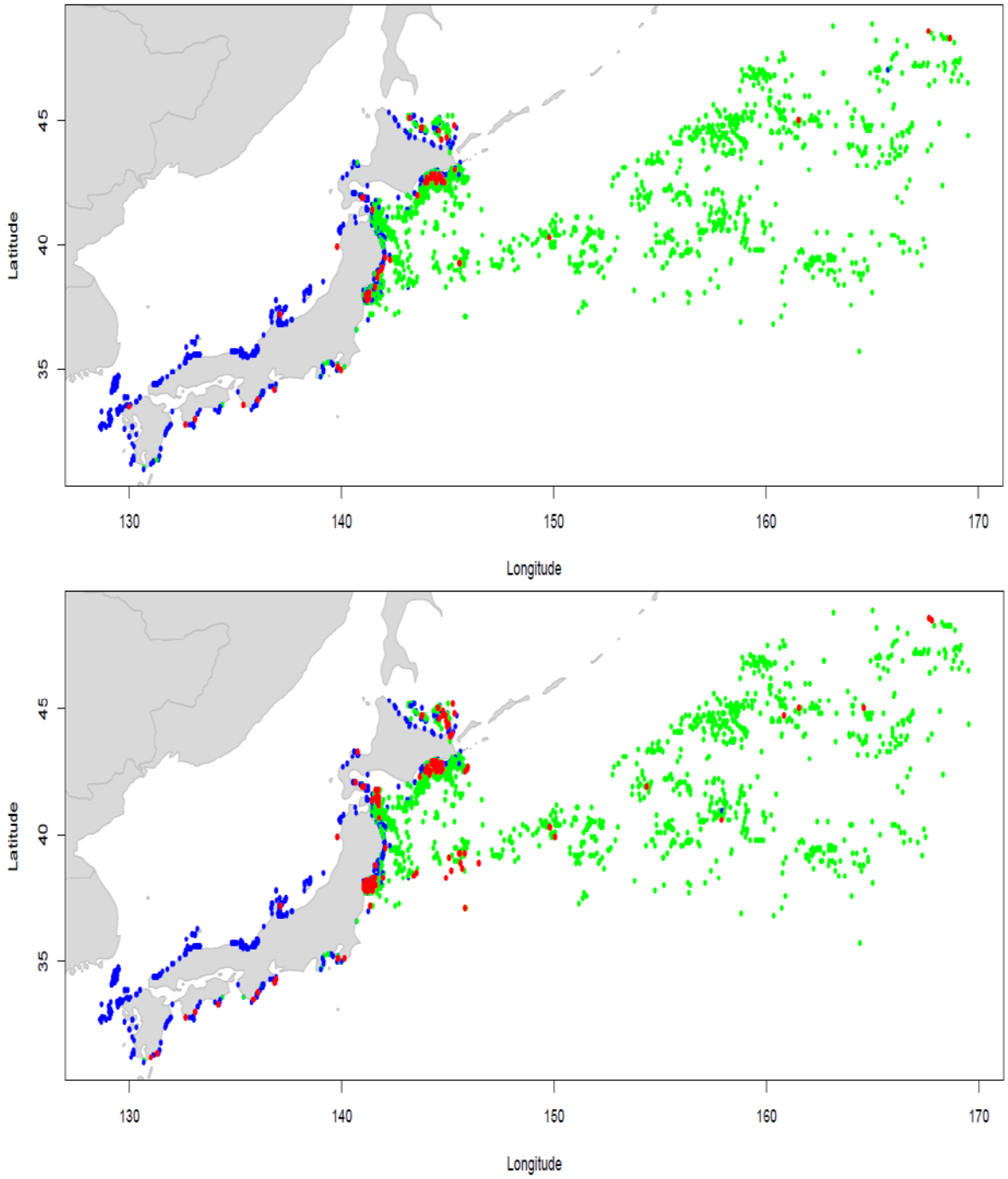
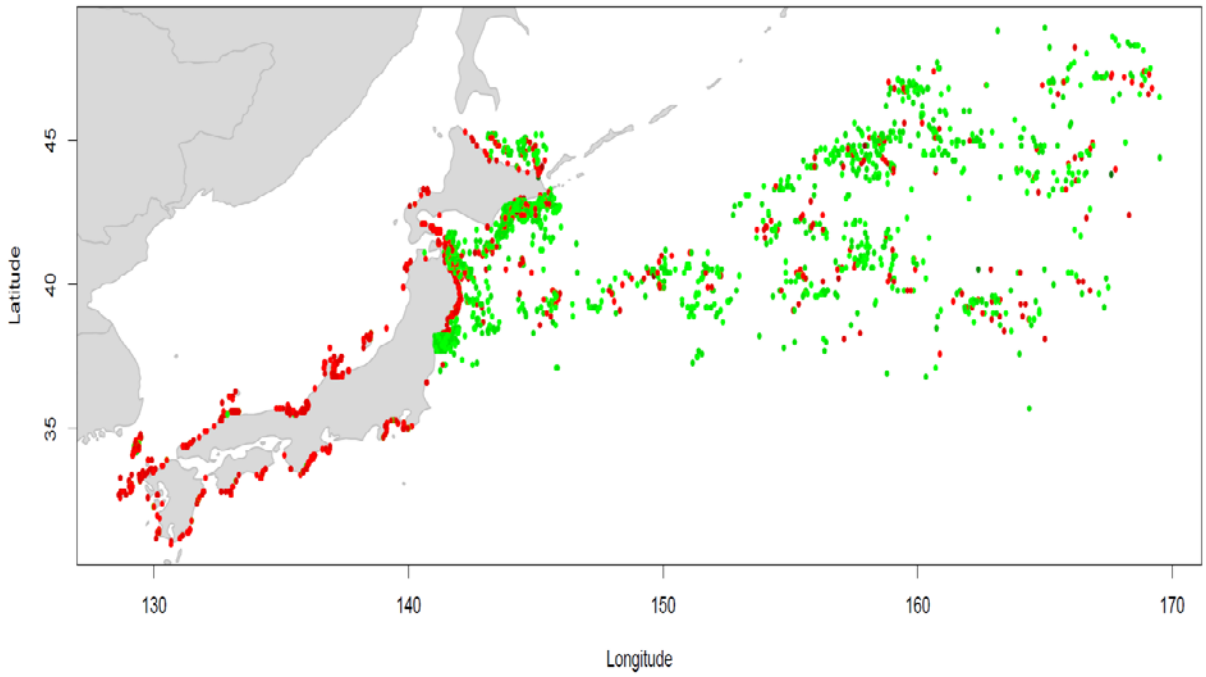


Figure 13: Output from BAPS when K=2.



Appendix 1:

R-command lines:

Geneland:

```
nloci       <- 16
mydata      <- read.table("DataSet_NPminke_WS2018_FINAL.txt",header=TRUE)
geno        <- mydata[,c(2:(2*nloci+1))]
nsamples    <- nrow(geno)
mycoord     <-  mydata[,c("longitude","latitude")]
MCMC(geno.dip.codom=geno,coordinates=mycoord,varnpop=TRUE,spatial=TRUE,freq.m
odel="Correlated",nit=1000000,thinning=100,rate.max=nsamples,npopmax=7,nb.nuclei.ma
x=3*nsamples,delta.coord=0)
PostProcessChain(nxdom=100,nydom=100,burnin=2000)
```

Illustrations:

To convert qmatrix values into posterior probabilities we executed the following R commands:

```
baps        <- read.table("baps.loglikelihood.txt")
baps2       <- exp(baps)
mysums      <- rowSums(baps2)
nrows       <- nrow(baps)
ncols       <- ncol(baps)
baymat      <- matrix(NA,nrows,ncols)
for (i in c(1:nrows))
{
        for(j in c(1:ncols))
{
baymat[i,j]     <- baps2[i,j]/mysums[i]
}
}
```

All plots were created in R using the getMap function of the 'rworldmap' package, by running the following commands:

```
plotdata            <- read.table("proba.pop.membership.indiv.txt")
colnames(plotdata)  <- c("longitude","latitude","c1","c2","c3","c4","c5","c6","c7")
xrange              <- c(min(plotdata$longitude),max(plotdata$longitude))
yrange              <- c(min(plotdata$latitude),max(min(plotdata$latitude))
library(rworldmap)
newmap              <-  getMap(resolution = "low")
plot(newmap, xlim=xrange, ylim=yrange)
points(plotdata$longitude,plotdata$latitude,pch=15,cex=1.5,col=plotdata$colour)
```

The plotdata$colour column, which indicates cluster assignment, was created as follows: First, we assigned colours to each cluster. Subsequently, individuals were assigned the colour of the cluster for which the individual had the highest posterior probability. We optionally

varied colour brightness (varying between black and the colour of the cluster) depending on the certainty of assignment, using the colorRampPalette function of the R package dichromat.

Allele frequency distributions:

Barplots of allele frequencies were created using the inbuilt R function barplot.
The correlation tests between cluster specific allele frequencies were executed with inbuilt R functions:

cor(freq1,freq2,method="pearson")
cor(freq1,freq2,method="spearman")

freq1 and freq2 are vectors of allele frequencies.