Fragmented habitat compensates for the adverse effects of genetic bottleneck

Ari Löytynoja[1,5], Pasi Rastas[1], Mia Valtonen[1,4], Juhana Kammonen[1], Liisa Holm[1,2], Lars Paulin[1], Jukka Jernvall[1,3], and Petri Auvinen[1]

[1]Institute of Biotechnology, HiLIFE, University of Helsinki, Finland
[2]Organismal and Evolutionary Biology Research Program, Faculty of Biosciences, University of Helsinki, Finland
[3]Department of Geosciences and Geography, Faculty of Science, University of Helsinki, Finland
[4]Current address: Natural Resources Institute Finland (LUKE), Helsinki, Finland
[5]Corresponding author: ari.loytynoja@helsinki.fi

Abstract

According to the population genetics theory, population subdivision helps organisms retain genetic diversity, crucial for adaptation in a changing environment. Habitat shape is thought to be important for generating and maintaining population subdivision, but empirical cases are needed to test this assumption. We studied Saimaa ringed seals, land-locked in a labyrinthine lake and recovering from a drastic bottleneck, by whole-genome sequencing 105 individuals and additional individuals from three other ringed seal subspecies. We analyzed the distribution of variation and the genetic relatedness among the individuals in relation to the habitat shape. Despite a severe history of a genetic bottleneck with prevalent homozygosity in Saimaa ringed seals, we found evidence for population structure mirroring the subregions of the lake. Highlighting the significance of habitat connectivity in conservation biology and the power of genomic tools in understanding its impact, the genome wide analyses showed the subpopulations to have retained unique variation and largely complementary patterns of homozygosity, giving hope to this endangered species.

Introduction

Conservation efforts aim at retaining and increasing the census size (N) of endangered species, but the significance of N_e, the effective population size, is also widely acknowledged (1–3). The concept of N_e was developed by Wright (4) and Fisher (5) and refers to the size of an idealized population that behaves genetically similarly to the real-life population under study. In conservation biology, the significance of N_e is that it determines the populations' ability to respond to evolutionary forces, e.g. to adapt to a changing environment or to survive under negative selection (6). While N_e can be affected by sex ratio, age structure and differences in reproductive success, one of its key determinants is the population structure. Wright (4; 7) showed that the metapopulation effective size, N_{eT}, is a function of N_T, the total population size, and F_{ST}, a differentiation among subpopulations, $N_{eT} \approx \frac{N_T}{1-F_{ST}}$. While this suggests that the effective population size is increased in structured populations, Wright made strong underlying assumptions about the nature of the subpopulations, and under realistic scenarios population subdivision is thought to reduce rather than increase the global effective size (6). Many results of Wright and Fisher's early work have later been derived from coalescence theory (8). There, N_e is defined as the coalescence time of sequences and is always increased by population subdivision.

Saimaa ringed seals (*Pusa hispida saimensis*) are a land-locked population of ringed seals in Lake Saimaa, Finland, and along with the Arctic (*P. h. hispida*), Okhotsk (*P. h. ochotensis*), Baltic (*P. h. botnica*), and Ladoga (*P. h. ladogensis*) ringed seals, one of the five recognized subspecies of the ringed seal. According to estimates (9), the census sizes of Okhotsk and Arctic ringed seals are counted in hundreds of thousands and millions, while the census size of the Baltic population is around 10,000 and that of the Ladoga population 3,000-5,000 individuals. While the population size of Saimaa ringed seals has recovered from 100-160 individuals in the early 1980s to more than 400 individuals in 2020 (10), the consequences of the extreme bottleneck to the population's adaptive potential and long-term prospects are unclear. Lake Saimaa is highly labyrinthine and has a shore length of 13,700 km for an area of 4,400 km^2, and further fragmented by ca. 14,000 islands. This is in stark contrast to neighboring Lake Ladoga, housing the only other freshwater subspecies of ringed seals, with a shore length 1,570 km for an area of 17,700 km^2. Ringed seals from different parts of Lake Saimaa have been shown to be genetically differentiated (11; 12), but the magnitude of this population stratification and its impact on the amount of genetic diversity and the subspecies' adaptive

potential has not been assessed in detail. Given the highly labyrinthine form of the lake, the Saimaa ringed seals provide a unique opportunity to empirically study the effects of population substructure on the preservation of genetic diversity and to learn about the significance of the environment shape for the conservation efforts of endangered species in general.

Results

Since 1980s, tissue samples from Saimaa ringed seals found dead have been collected for research needs in tissue banks maintained by Metsähallitus (the Finnish forest administration) and the University of Eastern Finland. To study the genetic variation in the Saimaa ringed seal population, we whole-genome sequenced 19 samples at medium sequencing coverage (mean 11.2X) and additional 86 samples at low coverage (mean 5.2X; Table S1). To place the findings into a context, we additionally sequenced at medium coverage (mean 10.9X) 9-12 samples from three other ringed seal subspecies, Ladoga, Baltic and Arctic (Fig. 1B; Table 1; Table S1). While the distributions of Baltic, Ladoga and Saimaa subspecies are geographically limited, it is noteworthy that all our Arctic samples come from Svalbard and may not represent the full variation of the subspecies' circumpolar distribution. This has a minor impact on analyses of the Saimaa subspecies and population and, for simplicity, we refer to the Svalbard population as the Arctic subspecies.

We used the 2.350 Gbp contig assembly for the Saimaa ringed seal and in most analyses focused on the 55.9% of the genome defined callable and unique by SNPable (13) and RepeatMasker (14); a subset of analyses were performed on the 100 longest contigs, the shortest of these being 5.48 Mbp and them in total covering 804 Mbp (34%) of the genome. Using bwa mem (15) and samtools (16) for read mapping, ANGSD (17) and beagle (18) for genotyping and phasing, we separately processed the Saimaa samples ('Saimaa data'; n=105) and the medium coverage samples from the four subspecies ('subspecies data'; n=50). Among the 1.313 Mbp sites, we found in total 5,894,190 single nucleotide polymorphisms (SNPs) in the Saimaa data and 10,163,875 SNPs in the subspecies data. We randomly selected two individuals from each non Saimaa subspecies and three individuals from Saimaa for a phylogenetic analysis, using a grey seal (*Halichoerus grypus*) individual as the outgroup ('phylogenetic data'; n=10). Starting from a large, unphased variant dataset genotyped with GATK (19), we extracted the samples and converted the variant information to DNA sequences by randomly picking one of the alleles.

Table 1: Number of shared and private SNPs (MAF 2.5%) and fixed differences in the four ringed seal populations.

	Arctic	Baltic	Ladoga	Saimaa
samples	10	12	9	19
total SNPs	8,387,034	8,987,504	8,244,735	5,185,101
private SNPs	38,634	61,299	40,540	321,482
private fixed	33	14	28	8,852

Saimaa subspecies is genetically diverged from others

We started by inferring the relationships between the four subspecies. We split the phylogenetic data into 100 non-overlapping 5 Mbp regions and inferred a phylogenetic tree for each region as well as for the concatenated data with RAxML (20), correcting for the number and base composition of non-variant sites. The maximum likelihood tree for the full data places Saimaa ringed seals at the basal position with 98% of the sequence trees supporting their monophyly (Fig. 1A). The evolutionary history of population samples is rarely tree-like, however, and many regional trees are incongruent with the main topology: apart from the monophylies of Saimaa and Ladoga, all internal branches have support values below 40%.

A consistent split is seen in the larger subspecies data. Using a minor allele frequency (MAF) of 2.5%, the four subspecies data contain 10,132,155 SNPs. Despite the smaller sample sizes (n=9-12), the Arctic, Baltic and Ladoga subspecies show significantly higher amounts of variation, from 8.24×10^6 to 8.99×10^6 SNPs each, than the Saimaa samples (n=19), in total 5.19×10^6 SNPs (Table 1). While 32.0% of the SNPs are shared among all subspecies, an additional 34.8% of the variation is shared among the three non-Saimaa subspecies (Fig. S1). In line with that, the Saimaa samples contain over 321,000 private SNPs and 8,852 fixed differences while the three other subspecies have 38,600-61,300 private SNPs and 14-33 fixed differences (Table 1, Fig. S1). Given the SNP sharing statistics, it is not surprising that the first PC of the principal component analysis (PCA) explains 21.8% of total genetic variation and separates the Saimaa samples from all others; the second PC explains 4.2% and separates the three other subspecies (Fig. 1B).

The PCA and SNP sharing statistics indicate that the Saimaa population has diverged from other ringed seals and lost genetic diversity in comparison to other subspecies. To understand this in more detail, we inferred the demographic histories and the subspecies split times with MSMC (21). We used the mutation rate estimated for polar bears, 1.826×10^{-8} substitutions per site per generation (22), and generation time of 10 years (23), and sampled ten two-individual sets from each population for within-population (in total two individuals) and between-populations (in total four individuals) anal-

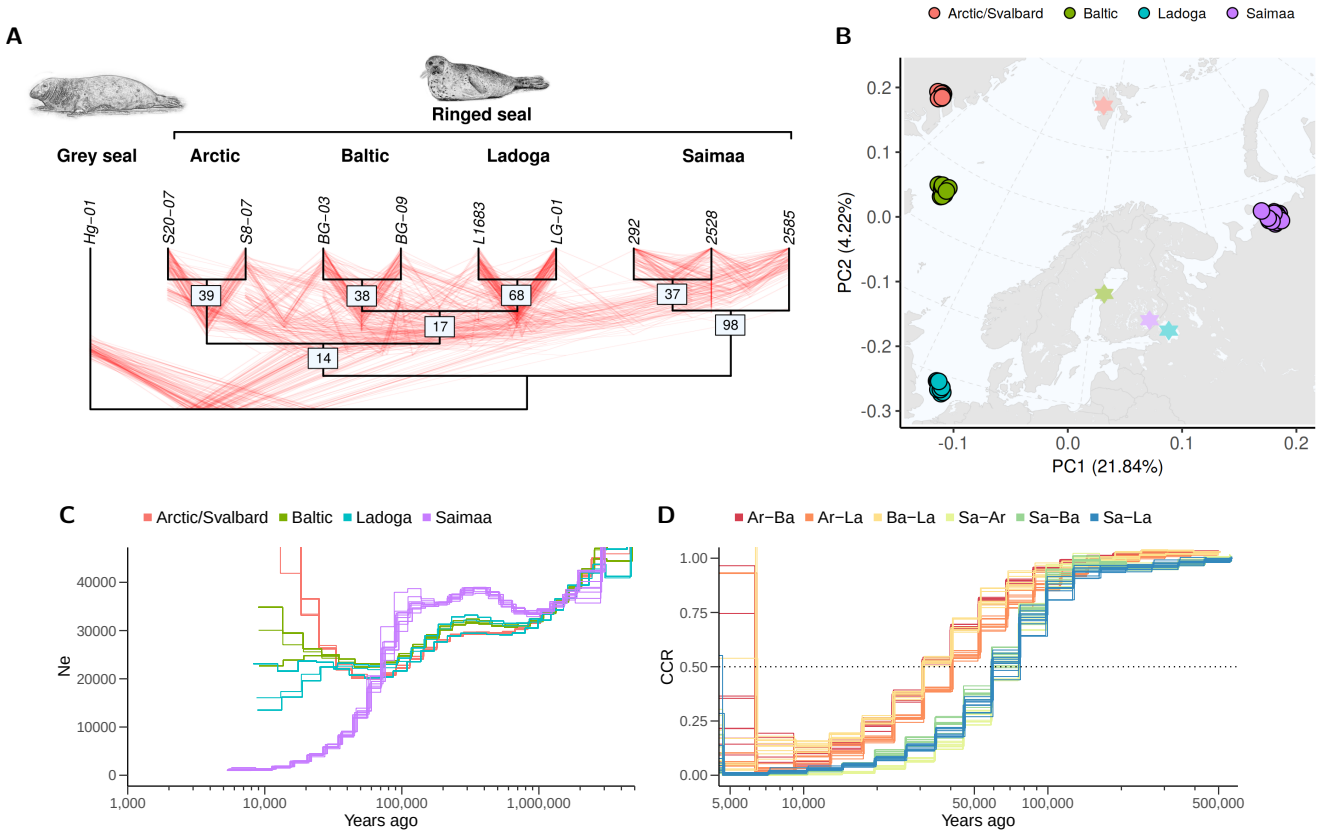

Fig. 1: Genetic divergence and demographic histories of the four ringed seal subspecies. (A) The maximum likelihood topology places Saimaa ringed seals at the basal position, but the 100 local trees (red; scaled to match) show incongruence and the support for most internal branches is low. (B) The plot shows the first two axes of the PCA of genetic variation in the four subspecies, indicated by colored dots. The background map shows the origin of the sampled subspecies, indicated by colored stars. (C) The lines indicate inferred ancestral population sizes (N_e) across time. For each subspecies, indicated by the different colors, the analysis was performed for ten randomly chosen pairs of two samples. (D) The lines indicate the cross population coalescence rate (CCR) for ten randomly chosen sets of four samples, two from each subspecies. The relative rate below one indicates (partial) population split, the rate of zero meaning complete isolation.

Table 2: Heterozygosity of individuals (H), nucleotide diversity within populations (π), genetic divergence between populations (d_{xy}; above diagonal) and population differentiation (F_{ST}; below diagonal) across the whole genome for the four ringed seal subspecies.

		Arctic	Baltic	Ladoga	Saimaa
	H	0.00207	0.00212	0.00205	0.00120
	π	0.00205	0.00205	0.00206	0.00127
d_{xy} or F_{ST}	Arctic	–	0.00214	0.00215	0.00217
	Baltic	0.0377	–	0.00216	0.00219
	Ladoga	0.0665	0.0491	–	0.00220
	Saimaa	0.3343	0.3185	0.3436	–

yses.

The N_e trajectories for Arctic, Baltic and Ladoga subspecies are very similar until 25,000 years ago (ya; Fig. 1C): after that, the Svalbard population grows strongly while the Baltic population increases slightly and the Ladoga population decreases in size. The trajectories of Saimaa individuals separate from others around 1,000,000 ya and show a clearly higher N_e until 100,000 ya. After that, the Saimaa N_e size dives and eventually drops below 1,000, more than an order of magnitude lower than for any other subspecies (Fig. 1C). The diverging estimates of recent N_e for Arctic, Baltic and Ladoga (Fig. 1C) seem to disagree with their very similar SNP statistics (Table 1, Fig. S1). This suggests that, due to the low sequencing coverage and 2.5% MAF cap in variant filtering, our SNP counts miss recent variants, affecting most the strongly growing Svalbard population. The MSMC method uses the lengths and ages of coalescent blocks (21) and is not significantly affected.

A growth in effective population size can be explained either by increase in actual census size or by population structure and admixture between distinct lineages (4; 7; 6; 8). In addition to estimation of N_e, MSMC allows studying the

cross population coalescence rate, i.e., the ratio of coalescence events between and within populations across the time. In that, a ratio of one indicates that the coalescence events happen equally frequently for samples coming from two current-day populations as for samples coming from one population; this implies that the two current-day populations were one at the corresponding time. Similarly, a ratio below one indicates that coalescence is more likely within one population and the two populations have been partially separated; a ratio of zero means that coalescence only happens within populations and the populations evolve independently.

Consistent with the N_e trajectories, the cross population coalescence rate ratios for Svalbard, Baltic and Ladoga pairs decrease below 0.5 around 30,000-40,000 ya (Fig. 1D); however, the values approach zero very recently indicating that the populations may not have fully separated. For pairs including Saimaa individuals, the cross population coalescence rate decreases below 0.5 around 60,000-85,000 ya, and the ratio is close to zero from 20,000 ya, indicating near complete separation of Saimaa from the other populations. While the results suggest some connection between Saimaa and other populations between 10,000 and 100,000 ya, the cross population coalescence rates involving Saimaa do not reach those of other populations before much deeper times (Fig. 1D). The results are consistent with Saimaa having a partially independent history and deriving a fraction of its ancestry from elsewhere. Similar signals are seen in other aquatic organisms in the Baltic Sea region, and e.g., some salmon stocks are thought to descend from eastern freshwater refugia (24).

Saimaa seals have lost much of their genetic variation

The SNP statistics are affected by the unequal sample sizes and to formally quantify the genetic variation we computed heterozygosity of individuals (H), nucleotide diversity within populations (π) and genetic divergence between populations (d_{xy}) for the different subspecies. The results confirm the lower variation in the Saimaa subspecies, with mean H and π of 0.0012 and 0.0013 differences per site, respectively, being slightly over half of that of other subspecies (Table 2). On the other hand, the highest d_{xy} estimates are for pairs involving Saimaa, and the d_{xy} among the three other subspecies are hardly higher than the within-population π. While this is consistent with the earlier findings, the very similar diversity estimates for the non-Saimaa populations are surprising given the orders of magnitude greater population size of the Arctic seals. This is probably explained by all our samples coming from Svalbard and the low coverage sequencing not being able to capture rare variants; however, we cannot exclude the impact of SNP ascertainment and variant imputation. Indeed, analysis of unphased raw data with MAF cap applied within population gives the highest π estimates for Svalbard, followed by Baltic and Ladoga and clearly behind by Saimaa (Table S2). This has no impact on the big picture, though, and Saimaa is clearly the odd-one out.

To study the distribution of variation across the genome we replicated the analyses for 250 kbp genomic windows. In line with the genome-wide mean estimates, the distributions of H and π for the three non-Saimaa populations peak around 0.002 differences per site and are nearly indistinguishable for the different subspecies, some Ladoga samples only having a small bump near zero for H (Fig. 2A,B). The picture is very different for Saimaa and, while the distribution of π peaks around 0.0012 (Fig. 2B), H is concentrated close to zero with another gently sloping bump above 0.002 (Fig. 2A). This indicates that the genomes of individual Saimaa seals contain long segments with basically no variation, but the copies of these low-variation regions tend to differ among the seals. Worryingly, the distribution of π is strongly left-leaning and there are numerous 250 kbp genome windows where the whole Saimaa population contains no variation; such pattern is not seen in the other subspecies (Fig. 2B). Distributions of d_{xy} across the genome confirm the Saimaa subspecies to contain private variation (Fig. S2, Table 2).

The lack of variation in Saimaa population implicates inbreeding and formation of runs of homozygosity (ROH) in the genome, their number and characteristics being informative about the history as well as the future prospects of the population. We inferred ROHs (25) on the 100 longest contigs, considering all sites, the possible variant calling errors making the estimates conservative. In line with their largest N_e, the Arctic individuals have the shortest ROHs and less than 10% of the genomes inferred as ROHs (Fig. 2C). The Ladoga seals have a greater proportion of the genome covered by ROHs than the Baltic ones (Fig. 2C), though the maximum ROH lengths are similar and neither show signs of contemporary inbreeding (26). In contrast to this, between 30 and 45% of the genome of Saimaa individuals is inferred as ROHs, and especially the medium sized ROHs of 1×10^5 - 5×10^5 bp are frequent (Fig. 2C). Moreover, due to the discontinuity of the reference genome, the inferred ROH lengths exceeding 2 Mbp are likely underestimates.

The small Saimaa population is highly structured

The results indicate that the Saimaa population has a deep independent history and its size has remained small for a long time. However, the lake is extremely fragmented and earlier studies have shown that the seals from different parts of the lake are genetically diverged (11; 12). We set out to resolve

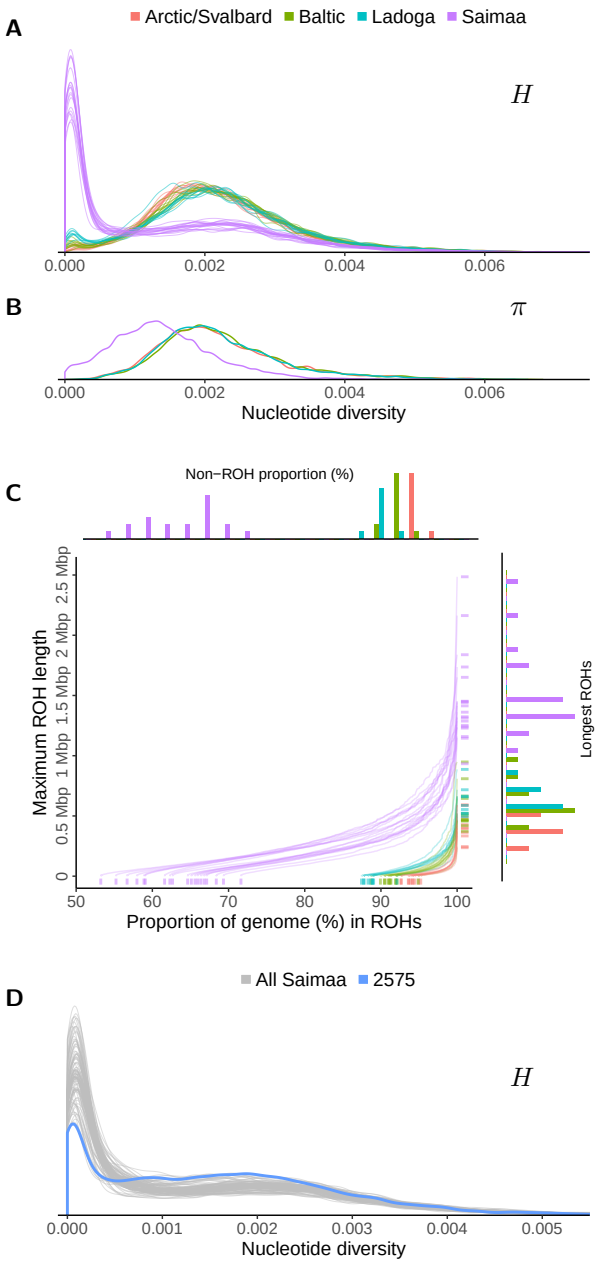

Fig. 2: Nucleotide diversity and runs of homozygosity across the genomes. The density lines show nucleotide diversity (A) for individuals (heterozygosity; H) and (B) for populations (π) in 250 kbp blocks; different subspecies are indicated with colors. In contrast to the three other subspecies, Saimaa individuals have many 250 kbp-genome regions with $H \approx 0$, i.e. parental copies contain basically no variation. While the distributions of H and π are nearly identical for the three other species, that of π for Saimaa shifts to the right and indicates that the loss of variation has hit different regions in different individuals and the population as a whole contains more variation. (C) The lines show the cumulative proportion of the genome (x axis) with the maximum ROH length below the given value (y axis). The histograms on top and right show the non-ROH proportion of the genome and the maximum ROH length, respectively. The lines on the left and bottom show the same for individual seals. (D) Heterozygosity of all 98 Saimaa samples with individual '2575' highlighted.

the population structure within the lake using 98 individuals with both the genetic and geographic information available.

In the PCA, the first two PCs explain 8.6 and 4.7% of the genetic variation and separate, respectively, the Northern Saimaa (NS) and the Southern Saimaa (SS) individuals from all others, leaving the Central Saimaa (CS) seals to connect the different clusters (Fig. 3A); additional PCs separate individuals from specific regions such as Kolovesi, the most fragmented and labyrinthine part of the lake (Fig. S3A-E). With one exception, the three groups found in the PCA occupy distinct lake regions connected by narrow straits (Fig. 3B). To test if the distribution of variation can be explained by simple isolation by distance in a highly fragmented landscape, we computed the water distances between the map locations and the genetic similarities between the samples. Two different measures of genetic similarities, identity by state (IBS; 27) distance and coancestry (28), correlate strongly (r=-0.543, $p < 2.2 \times 10^{-16}$) but the log-transform of the latter provides a better resolution and correlation over short geographic distances (r=-0.441, $p < 2.2 \times 10^{-16}$). However, fitting regression lines to full and partitioned data reveals that this correlation is largely caused by both genetic similarities and water distances being very different within and between the distinct lake regions (Fig. S4): the two show moderate correlation when considering only seals within a lake region (cor$=-0.200$, $p < 2.2 \times 10^{-16}$) but little for sample pairs coming from different lake regions (cor$=-0.070$, $p < 4.7 \times 10^{-8}$).

The coancestry matrix represents the estimated number of segments in the genomes of individuals that are most closely related to the corresponding parts in the genomes of other individuals, with high numbers of "chromosome painting" events indicating high genomic similarity. Similarly to the PCA analysis, the coancestry matrix contains three blocks of high relatedness, the corresponding individuals coming from the three distinct parts of the lake (Fig. 3C). However, the matrix also contains a few horizontal stripes, the darker color suggesting that these individuals share some genetic identity with most of the seals in the sample. Interestingly, one of these individuals, '2569', is the clearest deviation among the three clusters and is firmly amongst the NS samples in the PCA plot (Fig. 3A) despite being originally sampled deep within the CS region of the lake (Fig. 3B). In addition to being a migrant, '2569' shows the highest genomic heterozygosity (Table S3), consistent with ancestry of different backgrounds.

Our Saimaa data are of low sequencing coverage (Table S1) and imputation of missing genotypes could potentially create spurious signals of genetic identity. To assess the impact of population structure on the imputation process, we replicated the analyses with data imputed within subpopula-

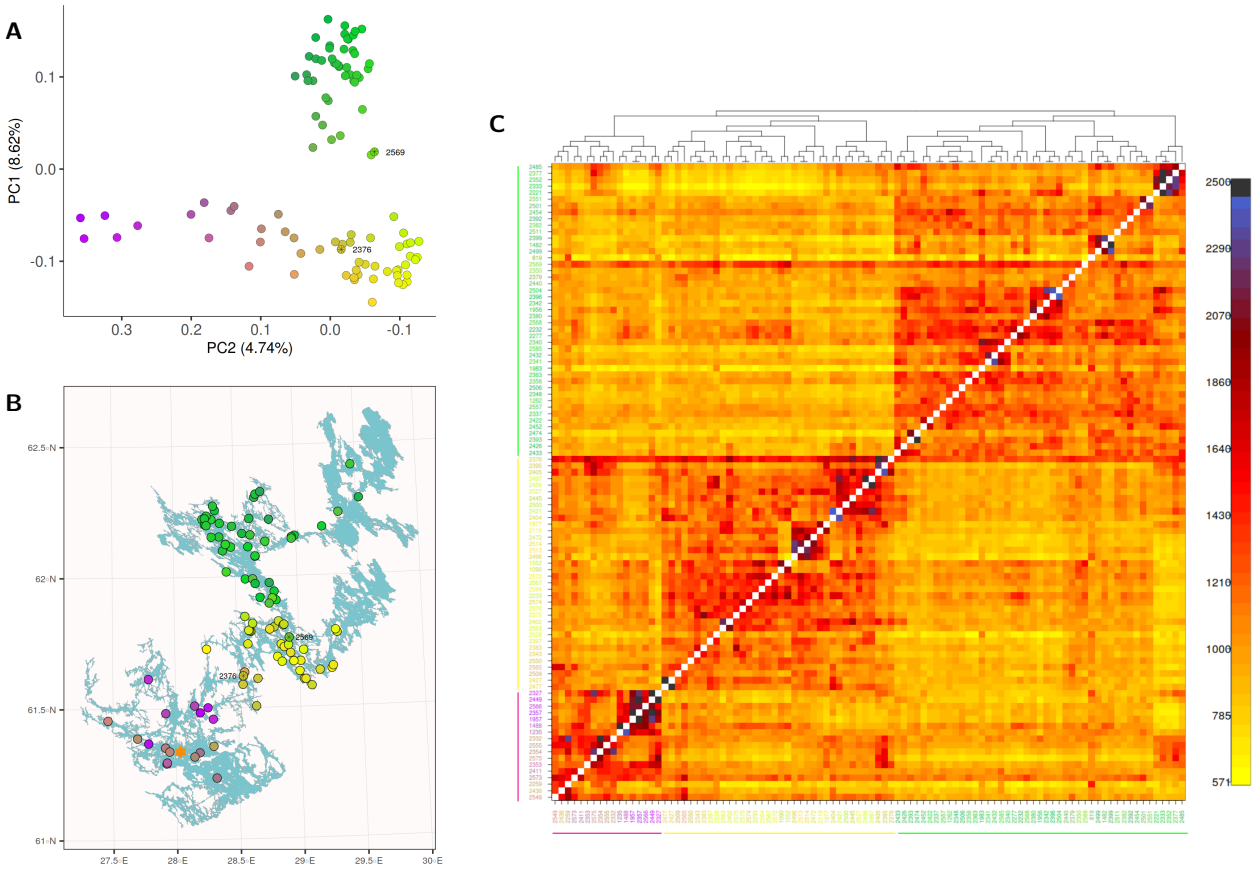

Fig. 3: Genetic variation within Saimaa. (A) The first two PCs explain 8.6 and 4.7% of the variation among the 98 samples. (A) The map of Saimaa shows the fragmentation and the labyrinthine shape of the lake. The sampling locations of the studied seals are indicated with dots matching the colors in the PCA plot. The orange star indicates the living area of a seal translocated in 1992. (C) The coancestry matrix contains three blocks of higher similarity (scale on the right with higher values indicating greater genetic similarity) that, with one exception, consist of samples from northern (top, indicated by green line), central (middle, yellow) and southern (bottom, magenta) part of the lake. The dendrogram indicates the inferred grouping of the samples, and the axis label colors (left, bottom) match those in the PCA plot. Samples '2569' and '2376' form the most noticeable horizontal stripes in the matrix and are indicated in the PCA plot and the map.

tions. Although the imputation groups stand out clearly with the partitioned data (Fig. S5), the stripes of higher coancestry are still visible, indicating that specific samples truly have mixed backgrounds; the impact of imputation strategy on the PCA is modest (Fig. S6). While neither option is perfect, we stick to the joint-imputed data in the subsequent analyses, the results potentially being conservative and the differences between the subpopulations slightly diluted.

Coancestry reveals a past translocation experiment?

In contrast to inferred migration events, one case of migration is empirically known as an adult female was translocated from the central Haukivesi (north-western part of Lake Saimaa) to the southern part of the lake in 1992 (29). No tissue sample was taken at the time, but the female is known to have reproduced in the area since 1994 (29) and was identified from the pelage pattern (30) to be still alive in 2020 (10). Interestingly, the PCA shows the seals around her living area to be genetically intermediate while the most distinctively "southern" seals are from the narrow entry through which the possible natural migration has to enter that part of the lake (Fig. 3A,B). There are 73 sample pairs with coancestry above 2000, all within the distinct parts of the lake (Fig. 3C, Fig. S7A) and the southern part having the highest proportion (Fig. S8A), but these do not show significant spatial clustering within the lake regions (Fig. S8B). On the other hand, there are 29 cross-region sample pairs with coancestry above 1500 (Fig. S7B) and these include the CS individuals '2569' and '2376' identified in the PCA and coancestry analyses (Fig. 3A,C). The latter is the most highly connected individual and one of only two samples showing high relatedness to all three parts of the lake (Fig. S7C). Interestingly, the other one is '2575' from the immediate vicinity of the

Table 3: Heterozygosity of individuals (H), nucleotide diversity within populations (π), genetic divergence between populations (d_{xy}; above diagonal) and population differentiation (F_{ST}; below diagonal) across the whole genome for the three Saimaa ringed seal subpopulations. Only the 22 samples born in the 2010s were included.

		Northern	Central	Southern
	H	0.00122	0.00119	0.00120
	π	0.00124	0.00116	0.00116
d_{xy} or F_{ST}	Northern	–	0.00125	0.00125
	Central	0.1140	–	0.00120
	Southern	0.0893	0.0862	–

translocated seal's living area, and in addition to links to CS individuals, the seal shows high similarity to one NS individual (Fig. S7D). While '2575' stands out among the SS samples, the coancestry analysis reveals a cluster of five NS seals (Fig. 3C, five columns on the right) that are more closely related to several SS seals, the highest values matching well with the translocated seal's living area (Fig. S3F).

Saimaa subpopulations contain private variation

Given the strong structuring and differentiation of the Saimaa population, it is of interest to quantify the private variation within the subpopulations and study its distribution across the genome. The sequenced samples originate from a long period of time and to learn about the current status, we extracted the 22 Saimaa samples born in the 2010s (Fig. S9), excluding '2569' as an apparent migrant (cf. Fig. 3) and two others due to atypically high homozygosity. The justification for the latter is that the samples predominantly consist of young pups found dead, and genetic diseases resulting from extensive inbreeding may have contributed to the deaths of the most homozygous individuals.

We first computed the estimates of genome-wide nucleotide diversity among the three parts of the lake and F_{ST}, a measure of population differentiation, between the subpopulations. Consistent with the coancestry estimates, the NS samples contain slightly more genetic variation than the CS and SS samples, both within individuals (H) and within populations (π) (Table 3). The estimates of d_{xy} are slightly higher than those of π indicating that subpopulations contain small amounts of private variation. The F_{ST} estimates between the subpopulations vary from 0.086-0.114 (Table 3), and are thus of the same magnitude as between human superpopulations (31). F_{ST} analysis is known to depend heavily on the ascertainment of the SNPs and the estimates from different datasets are not fully comparable (31). Nevertheless, in our analyses the Saimaa subpopulations (Table 3) show greater population differentiation than the three other ringed seal subspecies (Table 2). The F_{ST} estimates are not markedly different when all 98 Saimaa samples are included in the analysis, varying from 0.082-0.106.

As in the four subspecies analysis, the distributions of H for 250kbp windows concentrate close to zero (Fig. S10A) while those for π peak around 0.0012 (Fig. S10B). The bump around zero for π, meaning complete loss of variation within a window, is the greatest in SS individuals. Distributions of d_{xy} are slightly to the right of those of π and, crucially, the bump around zero is smaller (Fig. S10C), indicating that the subpopulations have lost variation in different genomic regions and contain private variation not present in other subpopulations. A closer look on the H distribution reveals one sample with a notably lower proportion of genomic segments with no variation. This turns out to be '2575' from the southern part of the lake with high coancestry with a northern individual (Fig. S7B). A reanalysis of the genomic distribution of H across all 98 Saimaa samples shows that '2575' is indeed unique and has the very lowest proportion of non-variable regions (Fig. 2D).

Saimaa subpopulations optimally preserve variation

Our analyses indicate that the shape of Lake Saimaa strongly affects the distribution of genetic variation among the local ringed seals and the subpopulations within the distinct parts of the lake have retained private genetic variation. On the other hand, the strong structuring of the population increases inbreeding and causes loss of variation both in individuals and within the subpopulations. To understand the benefits and drawbacks of strong population structure, we studied synthetic data. Using msprime (32), we simulated population histories approximating that of the Saimaa ringed seals (Fig. 4A,B, Fig. S11) and compared three different scenarios: 1) population decreases in size while remaining as one panmictic population; 2) population decreases in size after being split into N=2..5 subpopulations connected by star-like migration; and 3) population decreases in size after being split into N=2..5 fully isolated subpopulations (Fig. 4B). We assumed a founder population of 1,000 individuals 800 generations ago; a population split and start of decline 400 generations ago; and gradual reduction to 240 individuals; in the end, half of the individuals were sampled.

The results are in line with the expectations and show that with low or absent migration, the subpopulations lose much of the genetic variation; as the total final population size is fixed, the impact is greater the more subpopulations there are (Fig. 4C). The central population in the star-shaped structure retains more genetic variation than the peripheric population but the difference strongly depends on the number of sub-

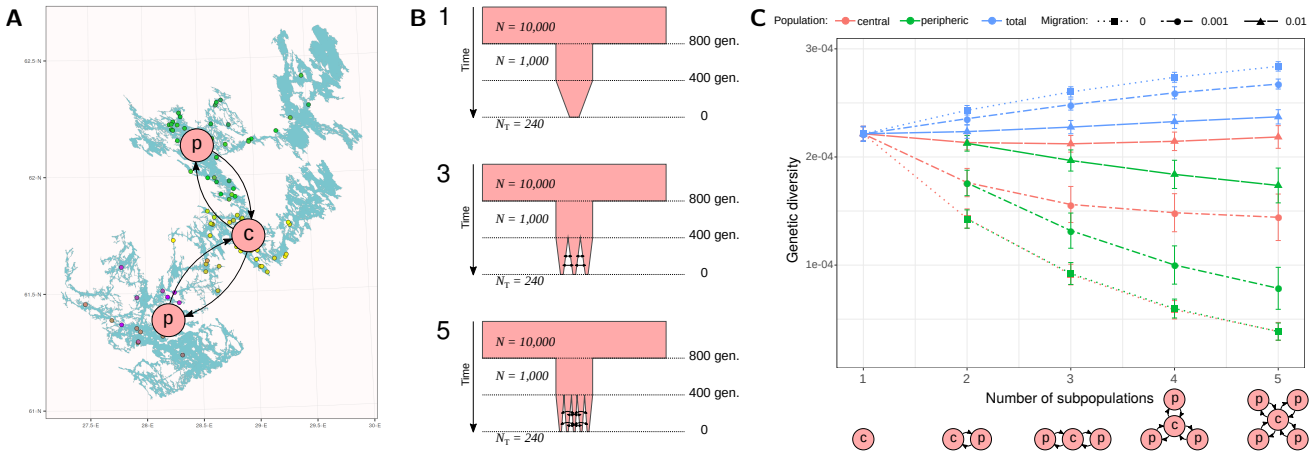

Fig. 4: Preservation of genetic variation in substructured populations. (A) The Saimaa ringed seal population shows strong differentiation with limited historical migration between the subpopulations. Due to the fragmented shape of the lake, the central subpopulation (c) connects the two peripheric (p) parts of the lake. (B) We simulated synthetic data for a population divided into N=1..5 subpopulations that were either connected with migration ($m > 0$) or evolving in isolation ($m = 0$). The population was founded by an initial population of 1,000 individuals 800 generations ago and this was then reduced to 240 individuals during the last 400 generations. (C) The plot shows the amount of nucleotide diversity (y axis) when a population of 240 individuals is either kept as one population (# subpopulations = 1) or divided into 2-5 subpopulations (x axis) and the migration rate varies (different line types); migration rate of 0 represents fully separated populations. The red, green and blue lines indicate the mean diversity over 100 replicates in the peripheric and central subpopulations, and in the total population, respectively.

populations and the amount of migration: with N=2, the central and the peripheric population are identical, and with a high migration rate their diversities differ little (Fig. 4C). In contrast to decreasing variation in the subpopulations, the amount of variation retained in the total population increases with the greater substructuring and the lower migration rate. Interestingly, if the migration rate is low (m=0.001), the three subpopulations (N=3) observed in Lake Saimaa seem to be close the sweet spot where the subpopulations lose one third (36.8%) of the variation retained by one panmictic population while gaining substantial increase (12.3%) in total variation due to substructuring (Fig. 4C). While the number of subpopulations above 3 does not affect the genetic divergence between the subpopulations (Fig. S12A), the resulting lower subpopulation size rapidly increases homozygosity of the individuals, especially those in the peripheric subpopulations (Fig. S12B).

Discussion

Lake Saimaa is unique in its fragmentation and labyrinthine outline. For a comparable area, the lake shoreline is roughly 58 times longer than if the lake would be a perfect circle. In contrast, Lake Ladoga has a relatively round shape with a shoreline only about 3.3 times longer than in a comparable circle. The labyrinthine shape of Lake Saimaa, together with its isolation, provides a natural setting to examine how population substructure responds to habitat fragmentation. Especially considering large mammals that typically require broad home ranges, Saimaa ringed seals offer a test case to measure the genetic effects of population subdivision.

Our analyses show that the Saimaa ringed seals are unique among the studied ringed seal subspecies and contain high numbers of private polymorphic sites and fixed nucleotide changes. The findings are consistent with the Saimaa population's longer isolation, or at least partial ancestry from an isolated source, and subsequent population bottleneck, as revealed by the demographic analysis. While the impact of the population bottleneck is evident in the total lack of variation in parts of the Saimaa genomes, a larger sample reveals a division into three strongly differentiated subpopulations. Importantly, the individuals and subpopulations have lost variation across different regions and, as a total, the Saimaa population retains some variation throughout most of the genome.

Detailed analyses show that the subpopulation divisions can be linked to narrow straits decreasing migration. In the analysis of genetic variation, the first principal component separates the northern subpopulation from the rest, the split coinciding with the narrow straits surrounded by the city Savonlinna. The second principal component separates the southern subpopulation, leaving the central part to connect the two extremities; other components separate geographically isolated individuals (cf. 12). While the subpopulation structure is suggestive of a star-like formation where peripheric parts of the lake are connected through a central population, the subpopulation in central Saimaa does not show the greater genetic variation as suggested by the simulation, possibly reflecting a lower historical population size.

The loss of genetic variation in Lake Saimaa seems drastic but the situation could be even worse. Our simulations examining different degree of migration between the subpopulations are suggestive that the Saimaa population is close to the sweet spot where the population as a whole is retaining or even gaining variation due to substructuring. Although the unique shape of Lake Saimaa and its impact on population structure and migration patterns are natural, one can draw parallels to extensive work on ecosystems and the significance of the number and connectivity of habitat patches (e.g. 33). We emphasize that our study should not be interpreted as an endorsement of artificial division of large habitats but rather a further example of the importance of migration and the value of even small subpopulations. Compared to previous reports on substructuring of the Saimaa population (12) and population analyses in general (34), our analyses also demonstrate that simple metrics such as F_{ST} alone provide limited information. Deciphering the genetic details underlying differentiation among subpopulations and individuals can reveal crucial knowledge for a species' conservation.

The Saimaa ringed seals seem to be a success story of conservation biology and the population size of the subspecies once listed as "Critically Endangered" (35) has grown from 100-160 individuals in the early 1980s to over 400 in 2020 (10). Despite the recent growth, the population is still highly vulnerable and especially the warmer winters and the shorter ice and snow cover pose a serious risk to breeding and early life of the pups (10). Nevertheless, our analyses indicate that the labyrinthine shape of Lake Saimaa has benefitted the retention of genetic diversity of the population as a whole, a result that can be useful in population management strategies that include translocations.

Materials and methods

Draft reference genome. The reference individual from Saimaa was sequenced with Pacific Biosciences RSII SMRT sequencing to 50X coverage, and assembled into 2,183 contigs using FALCON (v.0.3.0; 36). The contig assembly was further error-corrected by mapping 100X coverage of Illumina short-reads into the assembly and running the Pilon (v.1.16) improvement pipeline (37). A detailed description of the Saimaa ringed seal reference genome will be given in Kammonen et al. (in preparation).

Callable regions. Callable regions were defined with SNPable (13) and RepeatMasker (14). The SNPable mask file was generated using bwa (v.0.7.16; 15), k-mers of 35 bp and threshold of 50%. The fasta-formatted mask was converted to a bed-formatted positive mask using makeMappabilityMask.py from MSMC tools (21). The positive mask covered 86.7% of the genome. RepeatMasker (v.open-4.0.5) was run using rmblastn (v.2.6.0+) search engine and RepBase Update 20150807, and query species "Canis lupus familiaris". RepeatMasker masked 42.50% of the genome as repeats. After combining the two masks with bedtools, 55.9% of the genome was defined as callable and unique.

Population sequencing. All DNA sequencing was performed at the DNA Sequencing and Genomics Laboratory, Institute of Biotechnology, University of Helsinki. DNA was extracted with a NucleoMag Tissue kit (Macherey Nagel) and sequenced with NextSeq 500 and NovaSeq 6000 Illumina devices.

Variant calling. Bam files were constructed with bwa mem followed by GATK IndelRealigner (v.3.8-0; 19) and Picard Markduplicates (v.2.15.0; http://broadinstitute.github.io/picard). For the subspecies data and the Saimaa data, the variant calling was done with ANGSD (v.0.930; GATK genotyping mode: -GL 2; 17) and PCANGSD (v.0.981; number of PCA axes: -e 2 for the Saimaa dataset, -e 4 for the four subspecies dataset; 17) following the program documentation. The output of PCANGSD was converted to a vcf file using a custom script. Beagle (v.4.1 38) was run on the vcf file in its low coverage mode (gl=input.vcf.gz) and then in its genotype mode (gt=beagle_output.vcf.gz); for the Saimaa data, Beagle was run on all samples and for each subpopulation separately. Subsets of sites were extracted with bcftools (v.1.9-87; 16).

Phylogenetic analysis. The 'phylogenetic data' were extracted from a larger vcf file originally produced using the mapping protocol above followed by variant calling with GATK (v.3.7-0) and the program's best practises. The variants of the 100 longest contigs were converted to fasta files, each containing a 5 Mbp region separated by a 500 kbp spacer, using a custom script randomly picking one of the observed alleles or N for a missing genotype. For the concatenated alignment (consisting of 4,723,391 variable sites) and for each of the (in total 100) genomic fragments, a phylogenetic was inferred with RAxML (v.8.2.11; 20) using correction for missing data (-m ASC_GTRGAMMA --asc-corr=stamatakis). The density tree was generated with the R package phangorn (39).

Runs of homozygosity. ROHs were called with bcftools roh (v.1.9-87; 25) using the full data and assuming PL of 30 for the two least likely genotypes. The coverage and overlap of ROH segments of different length was computed with R package GenomicRanges (40). For the sum of ROHs, only contigs with length over 100 kbp were included, covering 98.9% of the total length of the assembly.

Nucleotide diversity and F_{ST}. Genome-wide heterozygosity per sample and genome-wide nucleotide diversity for the full population was computed using bcftools and custom

R code. Estimates were corrected for callable sites within windows of 250,000 bp, retaining windows with at least 60% of sites inferred as callable. Genome-wide mean heterozygosity and mean nucleotide diversity were computed across all callable sites. FST was computed with vcftools (v.0.1.13; 41) across all callable sites.

Coancestry, IBS distances and principal component analysis. Coancestry (chromosome sharing) was computed with fineSTRUCTURE (v.4.1.1; 28) using the full data for the 100 longest contigs. Identity-by-state (IBS) distances (1-ibs) were computed with Plink (v.1.90b6.9; 27) over the callable and unique sites. The principal component analysis (PCA) was performed with smartpca (v.16000) from the EIGENSOFT package (42) using the 100 longest contigs after MAF 2.5% filtering. The geographic data for the background map was from the R package rnaturalearth (https://github.com/ropensci/rnaturalearth).

Water distances. The Lake Saimaa map was downloaded from the Finnish Environment Institute open data service (http://paikkatieto.ymparisto.fi/lapio/latauspalvelu.html) and the locations of the seal carcasses were received from Natural Resources Institute Finland. The Saimaa map was converted to 576x717 raster (0.25km spacing), using R packages raster (https://rspatial.org/raster) and fasterize (https://github.com/ecohealthalliance/fasterize), and the shortest paths between the carcass locations were computed with R package gdistance (https://github.com/DistanceDevelopment/Distance), allowing eight movement directions.

Data simulation. The simulated ancestral population started with 10,000 individuals and was reduced to 1,000 individuals 800 generations ago; after 400 generations this population was split into N (N=1-5) subpopulations and, during the last 400 generations, reduced to 240 individuals; in the end, half of the individuals were sampled. We assumed mutation and recombination rates of 1×10^{-8} events per site per generation, migration rate (i.e. the fraction of population migrating on each generation) of 0, 1×10^{-3}, 1×10^{-2} and 1×10^{-1}, and simulated 100 replicates of 100 Mbp long sequence. In this setup, one "subpopulation" represents the panmictic scenario and migration rate of 0 the fully isolated subpopulations. From each simulation outcome we computed nucleotide diversity for the central and peripheric subpopulations and for the total population, the genetic divergence between central and peripheric subpopulations, and the proportion of the total segregating sites that are homozygous within different subpopulations. The data were simulated with msprime (v.0.7.4; 32).

References

1. Frankham, R. Effective population size/adult population size ratios in wildlife: a review. *Genet. Res.* **66**, 95–107 (1995).

2. Palstra, F. P. & Ruzzante, D. E. Genetic estimates of contemporary effective population size: what can they tell us about the importance of genetic stochasticity for wild population persistence? *Mol. Ecol.* **17**, 3428–3447 (2008).

3. Hare, M. P. *et al.* Understanding and estimating effective population size for practical application in marine species management. *Conserv. Biol.* **25**, 438–449 (2011).

4. Wright, S. Evolution in mendelian populations. *Genetics* **16**, 97–159 (1931).

5. Fisher, R. A. The distribution of gene ratios for rare mutations. *Proceedings of the Royal Society of Edinburgh* **50**, 204–219 (1931).

6. Waples, R. S. Spatial-temporal stratifications in natural populations and how they affect understanding and estimation of effective population size. *Mol. Ecol. Resour.* **10**, 785–796 (2010).

7. Wright, S. Size of population and breeding structure in relation to evolution. *Science* **87**, 430–431 (1938).

8. Wakeley, J. *Coalescent Theory: An Introduction* (W. H. Freeman, 2008), 1 edn.

9. NAMMCO. Ringed seals. https://nammco.no/topics/ringed-seal.

10. Kunnasranta, M. *et al.* Sealed in a lake — biology and conservation of the endangered Saimaa ringed seal: A review. *Biol. Conserv.* **253**, 108908 (2021).

11. Valtonen, M., Palo, J. U., Ruokonen, M., Kunnasranta, M. & Nyman, T. Spatial and temporal variation in genetic diversity of an endangered freshwater seal. *Conserv. Genet.* **13**, 1231–1245 (2012).

12. Valtonen, M. *et al.* Causes and consequences of fine-scale population structure in a critically endangered freshwater seal. *BMC Ecol.* **14**, 22 (2014).

13. Li, H. SNPable. http://lh3lh3.users.sourceforge.net/snpable.shtml. Accessed: 2021-10-20.

14. Smit, A., Hubley, R. & Green, P. RepeatMasker open-3.0 (1996).

15. Li, H. Aligning sequence reads, clone sequences and assembly contigs with bwa-mem (2013). http://arxiv.org/abs/1303.3997.

16. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10**, giab008 (2021).

17. Korneliussen, T., Albrechtsen, A. & Nielsen, R. ANGSD: Analysis of next generation sequencing data. *BMC Bioinformatics* **15**, 356 (2014).

18. Browning, B. L., Tian, X., Zhou, Y. & Browning, S. R. Fast two-stage phasing of large-scale sequence data. *Am. J. Hum. Genet.* **108**, 1880–1890 (2021).

19. McKenna, A. *et al.* The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).

20. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).

21. Schiffels, S. & Wang, K. MSMC and MSMC2: The multiple sequentially Markovian coalescent. *Methods Mol. Biol.* **2090**, 147–166 (2020).

22. Liu, S. *et al.* Population genomics reveal recent speciation and rapid evolutionary adaptation in polar bears. *Cell* **157**, 785–794 (2014).

23. Palo, J. U., Mäkinen, H. S., Helle, E., Stenman, O. & Väinölä, R. Microsatellite variation in ringed seals (*Phoca hispida*): genetic structure and history of the Baltic Sea population. *Heredity* **86**, 609–617 (2001).

24. Säisä, M. *et al.* Population genetic structure and postglacial colonization of Atlantic salmon (*Salmo salar*) in the Baltic Sea area based on microsatellite DNA variation. *Can. J. Fish. Aquat. Sci.* **62**, 1887–1904 (2005).

25. Narasimhan, V. *et al.* BCFtools/RoH: a hidden Markov model approach for detecting autozygosity from next-generation sequencing data. *Bioinformatics* **32**, 1749–1751 (2016).

26. Ceballos, F. C., Joshi, P. K., Clark, D. W., Ramsay, M. & Wilson, J. F. Runs of homozygosity: windows into population history and trait architecture. *Nat. Rev. Genet.* **19**, 220–234 (2018).

27. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).

28. Lawson, D. J., Hellenthal, G., Myers, S. & Falush, D. Inference of population structure using dense haplotype data. *PLoS Genet.* **8**, e1002453 (2012).

29. Koskela, J. T., Kunnasranta, M., Hämäläinen, E. & Hyvärinen, H. Movements and use of haul-out sites of radio-tagged Saimaa ringed seal (*Phoca hispida saimensis* Nordq.) during the open-water season. *Ann. Zool. Fennici* **39**, 59–67 (2002).

30. Chehrsimin, T. *et al.* Automatic individual identification of Saimaa ringed seals. *IET Comput. Vision* **12**, 146–152 (2018).

31. Bhatia, G., Patterson, N., Sankararaman, S. & Price, A. L. Estimating and interpreting FST: the impact of rare variants. *Genome Res.* **23**, 1514–1521 (2013).

32. Kelleher, J., Etheridge, A. M. & McVean, G. Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLoS Comput. Biol.* **12**, e1004842 (2016).

33. Wintle, B. A. *et al.* Global synthesis of conservation studies reveals the importance of small habitat patches for biodiversity. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 909–914 (2019).

34. Nyman, T. *et al.* Demographic histories and genetic diversities of Fennoscandian marine and landlocked ringed seal subspecies. *Ecol. Evol.* **4**, 3420–3434 (2014).

35. Kovacs, K. M. *et al.* Global threats to pinnipeds. *Mar. Mamm. Sci.* **28**, 414–436 (2012).

36. Chin, C.-S. *et al.* Phased diploid genome assembly with single-molecule real-time sequencing. *Nat. Methods* **13**, 1050–1054 (2016).

37. Walker, B. J. *et al.* Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* **9**, e112963 (2014).

38. Browning, B. L. & Browning, S. R. Genotype imputation with millions of reference samples. *Am. J. Hum. Genet.* **98**, 116–126 (2016).

39. Schliep, K. phangorn: phylogenetic analysis in R. *Bioinformatics* **27**, 592–593 (2011).

40. Lawrence, M. *et al.* Software for computing and annotating genomic ranges. *PLoS Comput. Biol.* **9**, e1003118 (2013).

41. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).

42. Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet.* **2**, e190 (2006).

Acknowledgements

We acknowledge UEF and Metsähallitus for the Saimaa samples, M. Verevkin for the Ladoga samples, M. Kunnasranta, P. Timonen, S. Oksanen, and J. Vierimaa for the Baltic samples, and H. Routti for the Svalbard samples; the personnel of the DNA sequencing and genomics laboratory for the DNA isolations and sequencing; and CSC – IT Center for Science Ltd. for computational resources. Claudius Kratochwil and Atte Moilanen are thanked for discussions and comments.

Funding

JJ and PA are funded by the Jane and Aatos Erkko Foundation (4-2013, 5-2017) and PA by the LIFE programme (LIFE19NAT/FI/000832).

Authors contributions

AL, PA and JJ conceived the paper. AL designed and performed the analyses, and wrote the first draft. PR imputed and genotyped the data. MV advised in analyses of Saimaa population structure. JK assembled the reference genome. LH contributed to the genome annotation. LP and PA coordinated the DNA sequencing. AL and JJ finalized the text. JJ and PA head the Saimaa ringed seal genome project. All authors commented the manuscript.

Competing interests

The authors declare no competing interests.

Data and code availability

The DNA sequencing data and the draft contig assembly have been submitted to ENA under accessions PRJEB51012 and PRJEB51013, respectively. The code for msprime simulation and computation of nucleotide diversity estimates are available at https://github.com/ariloytynoja/saimaa-population-structure.

Fragmented habitat compensates for the adverse effects of genetic bottleneck

Ari Löytynoja, Pasi Rastas, Mia Valtonen, Juhana Kammonen, Liisa Holm, Lars Paulin, Jukka Jernvall, and Petri Auvinen

Supplementary material

List of Figures

S1 SNP sharing among the four subspecies . . . 14
S2 Nucleotide diversity across the genome . . . 15
S3 Distribution of PCs and coancestry on the map . . . 16
S4 Correlation of coancestry and water distances. . . . 17
S5 Impact of imputation strategy on coancestry analysis . . . 18
S6 Impact of imputation strategy on PCA . . . 19
S7 Samples with very high coancestry estimates . . . 20
S8 Coancestry and water distances . . . 21
S9 Saimaa seals born in the 2010s . . . 22
S10 Nucleotide diversity across the genome . . . 23
S11 Simulation models . . . 24
S12 Genetic variation in substructured populations . . . 25

List of Tables

S1 Sequencing coverage . . . 26
S2 Effect of SNP ascertainment and phasing on π. . . . 27
S3 Metadata and per sample heterozygosity . . . 28
S4 Impact of imputation strategy . . . 29

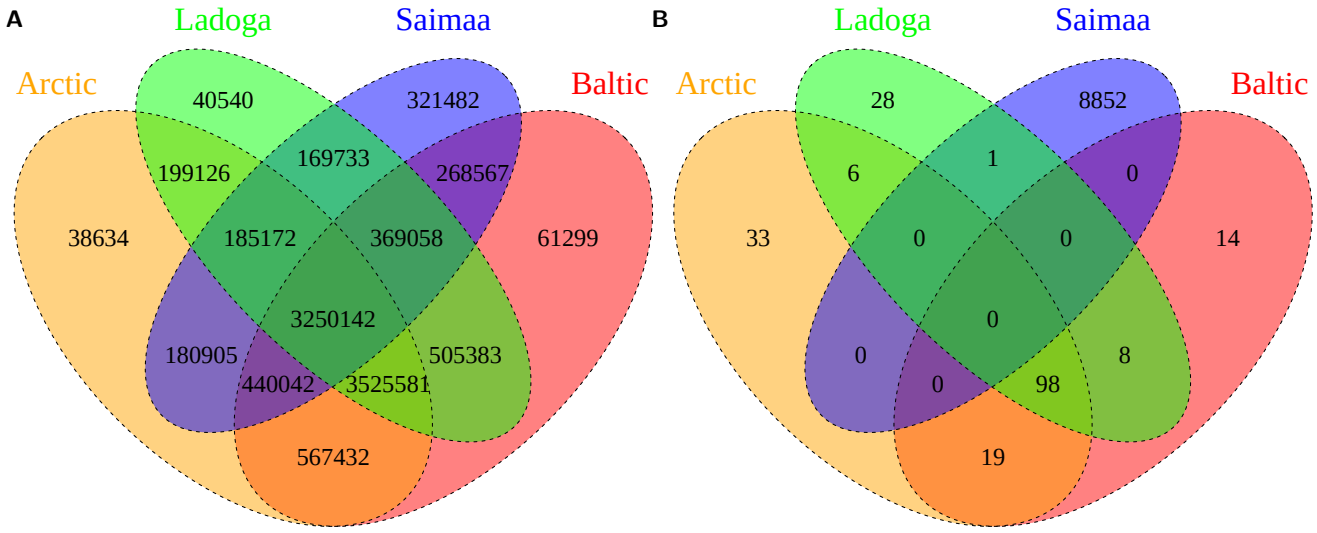

Fig. S1: SNP sharing among the four subspecies. (A) Sharing of the 10,123,096 polymorphic sites. (B) Sharing of the 9,059 fixed sites. Together the two sets make the 10,132,155 variable sites after MAF 2.5% filtering.

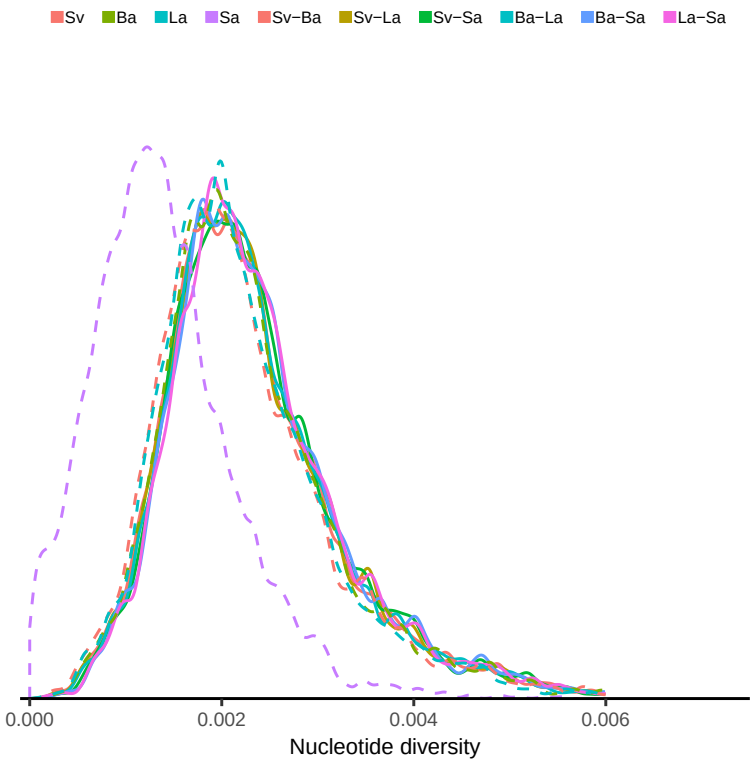

Fig. S2: Nucleotide diversity across the genome. The density lines show the mean nucleotide diversity (π; dashed) and the mean nucleotide divergence (d_{xy}; solid) in 250 kbp blocks for different subspecies (colors).

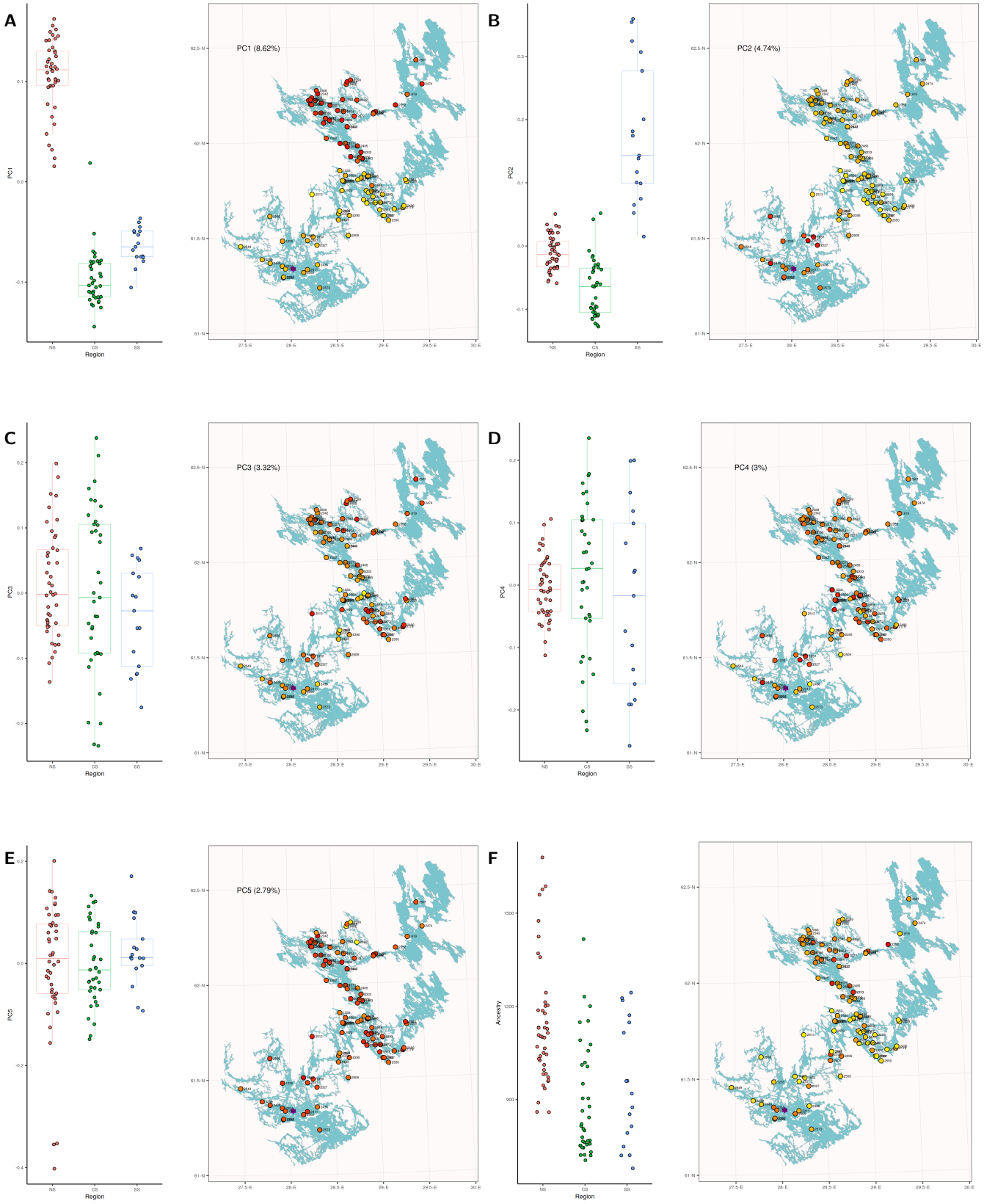

Fig. S3: Distribution of PCs and coancestry on the map. (A-E) The distribution of the first five PCs in different subpopulations (left) and on the map (right). (F) The mean coancestry with the five NS samples identified as a diverged clade in the fineSTRUCTURE analysis. The dots on the map are colored by the PC/coancestry value, from yellow (low) to red (high).

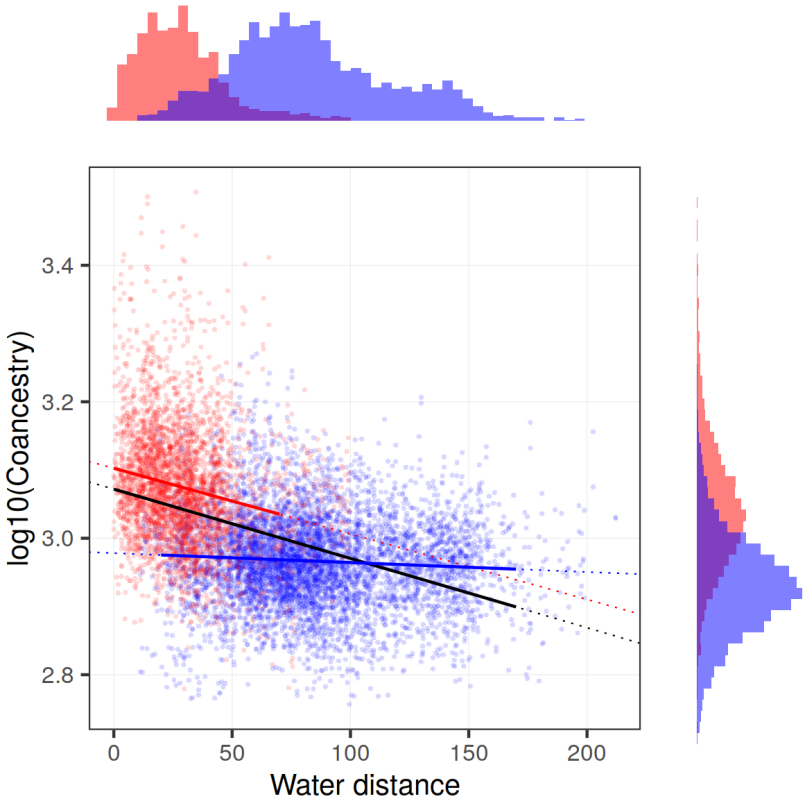

Fig. S4: Correlation of coancestry and water distances. Log10-transformed coancestry estimates and water distance between sample pairs correlate (cor=-0.441, $p < 2.2 \times 10^{-16}$) but this is largely explained by pairs coming either from the same or from different lake regions, indicated by red and blue points, respectively. Histograms on top and right show the distribution of pairwise water distances and pairwise log10-transformed coancestry estimates. The black, red and blue regression lines are fitted to full data, to red dots and to blue dots, respectively.

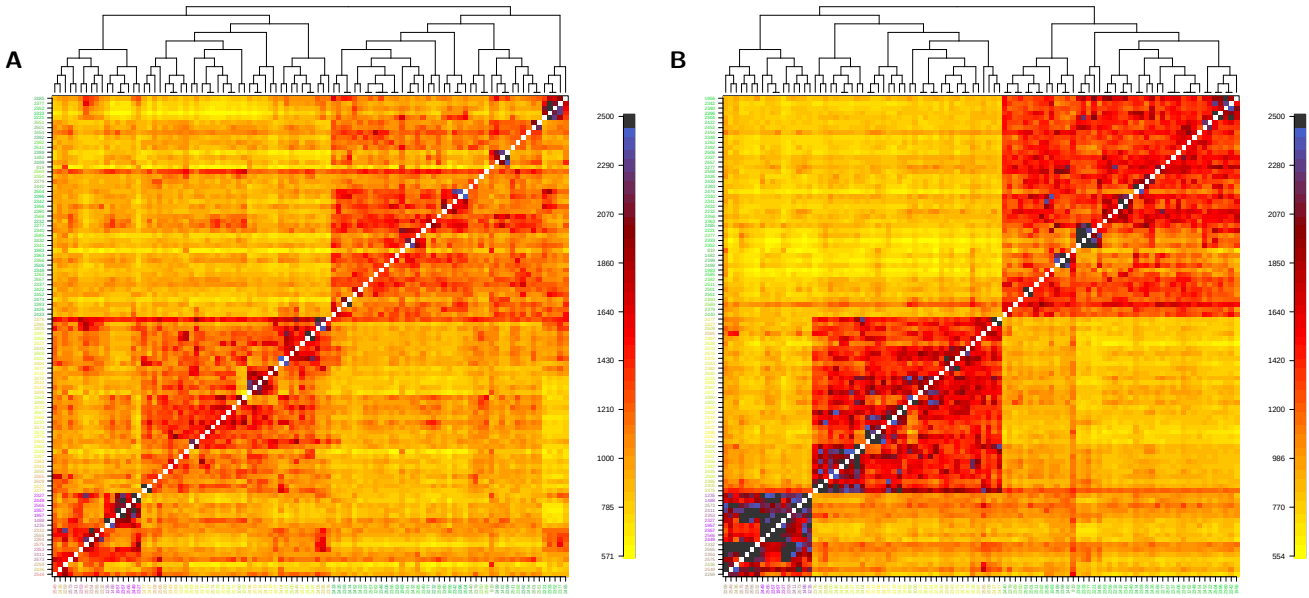

Fig. S5: Impact of imputation strategy on coancestry analysis. (A) Coancestry matrix inferred from data imputed as one batch. (B) Coancestry matrix inferred from data imputed in three separate batches. The samples are in different order in the two matrices.

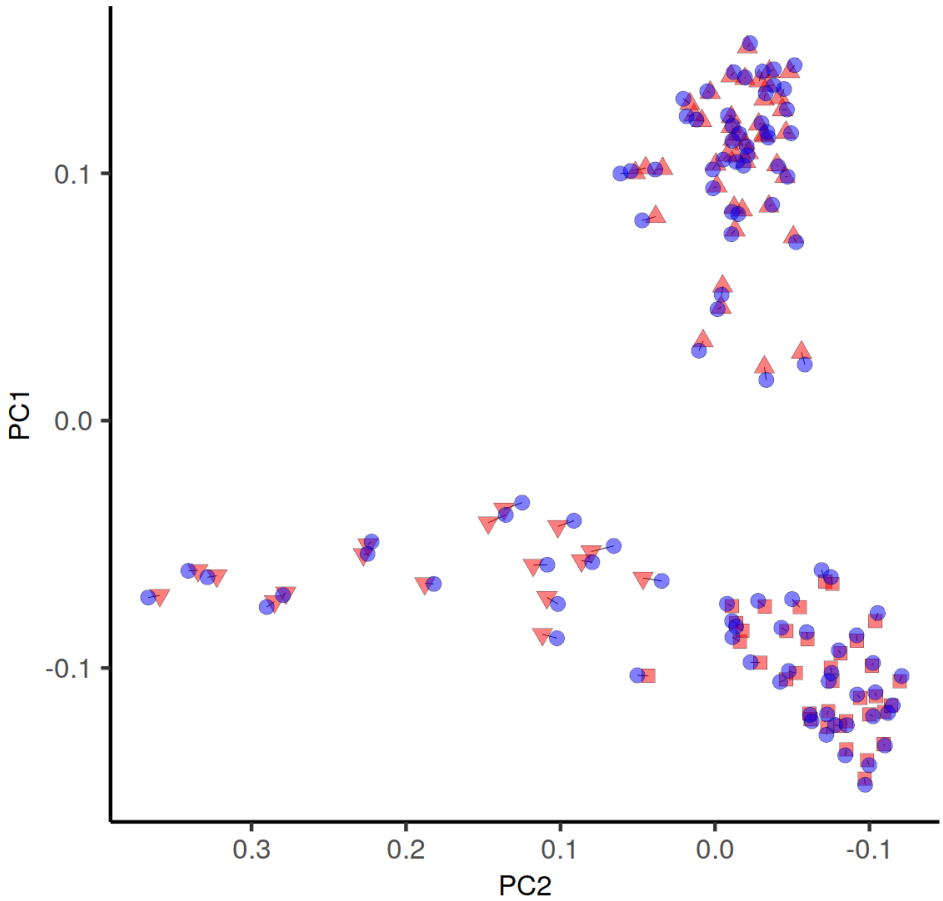

Fig. S6: Impact of imputation strategy on PCA. The impact of the imputation strategy is predictable but modest, the three clusters being more compact with separate imputation (red, batches indicated with different shapes) and the intermediate individuals moving slightly towards the non-native subpopulation with joint imputation (blue, connected with lines).

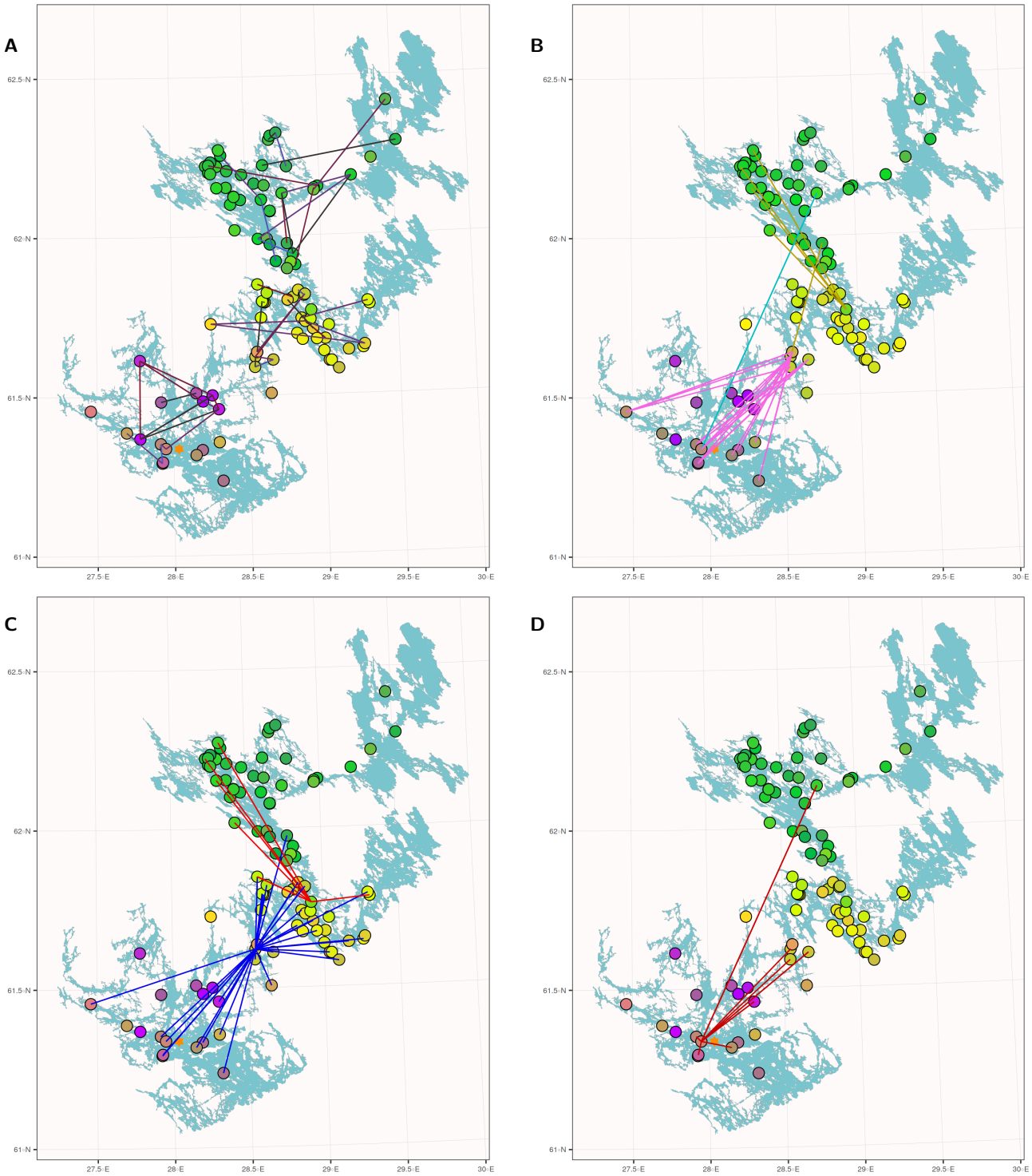

Fig. S7: Samples with very high coancestry estimates. (A) 25, 24 and 24 sample pairs, connected with lines, in the northern, central and southern part of Saimaa show coancestry above 2000. Line colors match the scale in Fig. 3d. (B) 29 sample pairs show coancestry above 1500 despite coming from different parts of the lake. Lines are colored by the subpopulations. (C) Sample pairs of '2569' (blue) and '2376' (red) with coancestry above 1500. (D) Sample pairs of '2575' with coancestry above 1500.

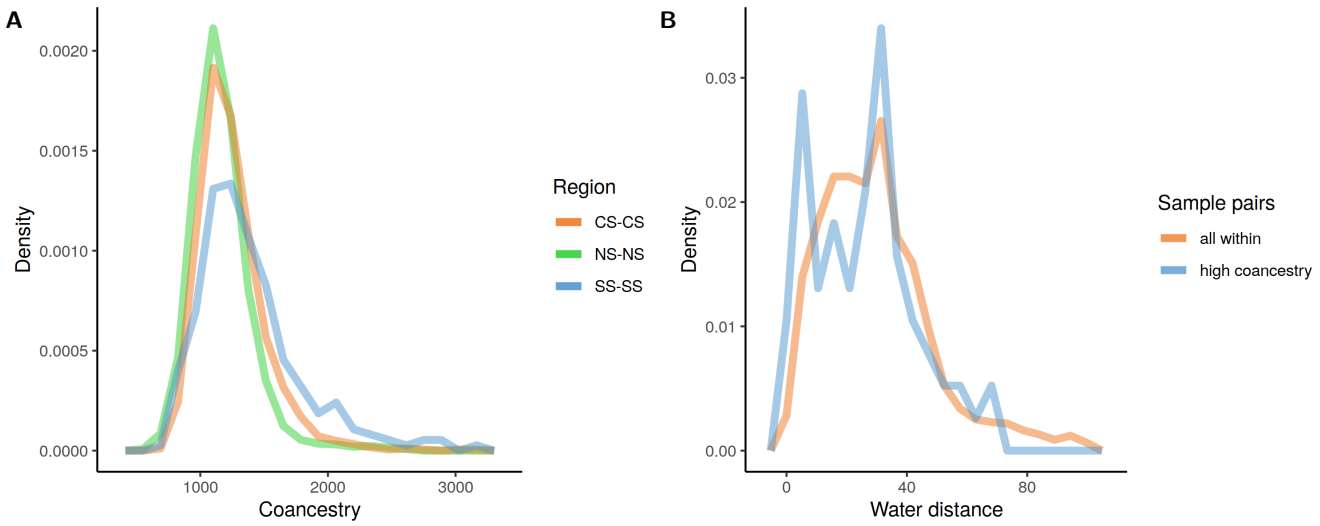

Fig. S8: Coancestry and water distances. (A) Distribution of coancestry within different regions. (B) Water distances between all sample pairs within the three regions and between samples showing coancestry above 2000.

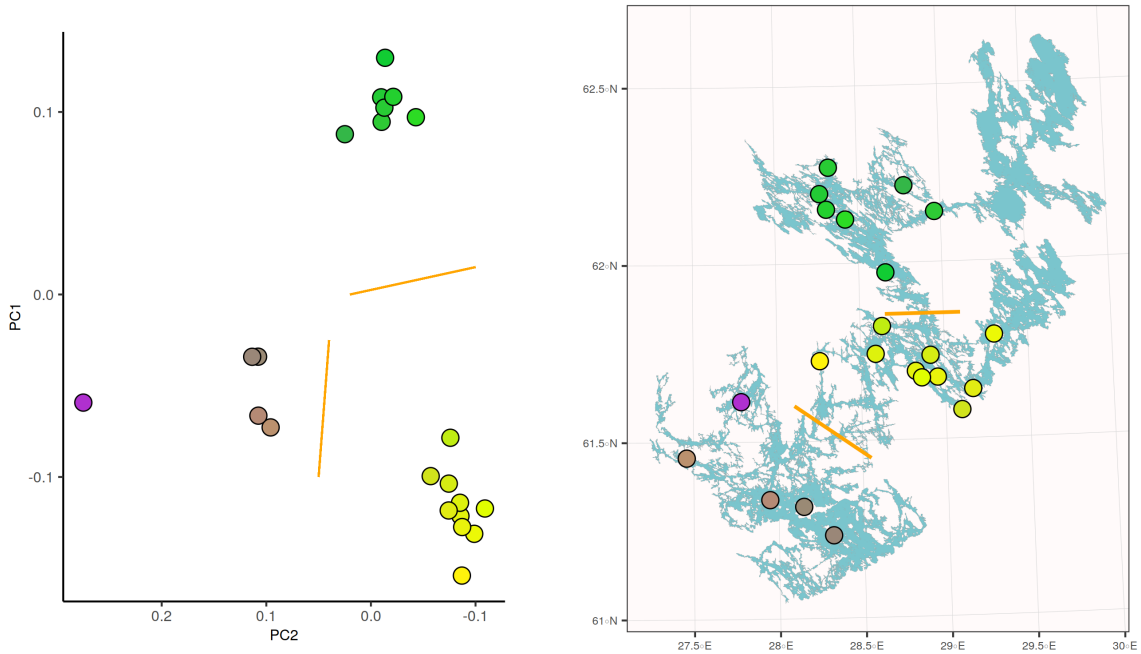

Fig. S9: Saimaa seals born in the 2010s chosen to represent the current population.

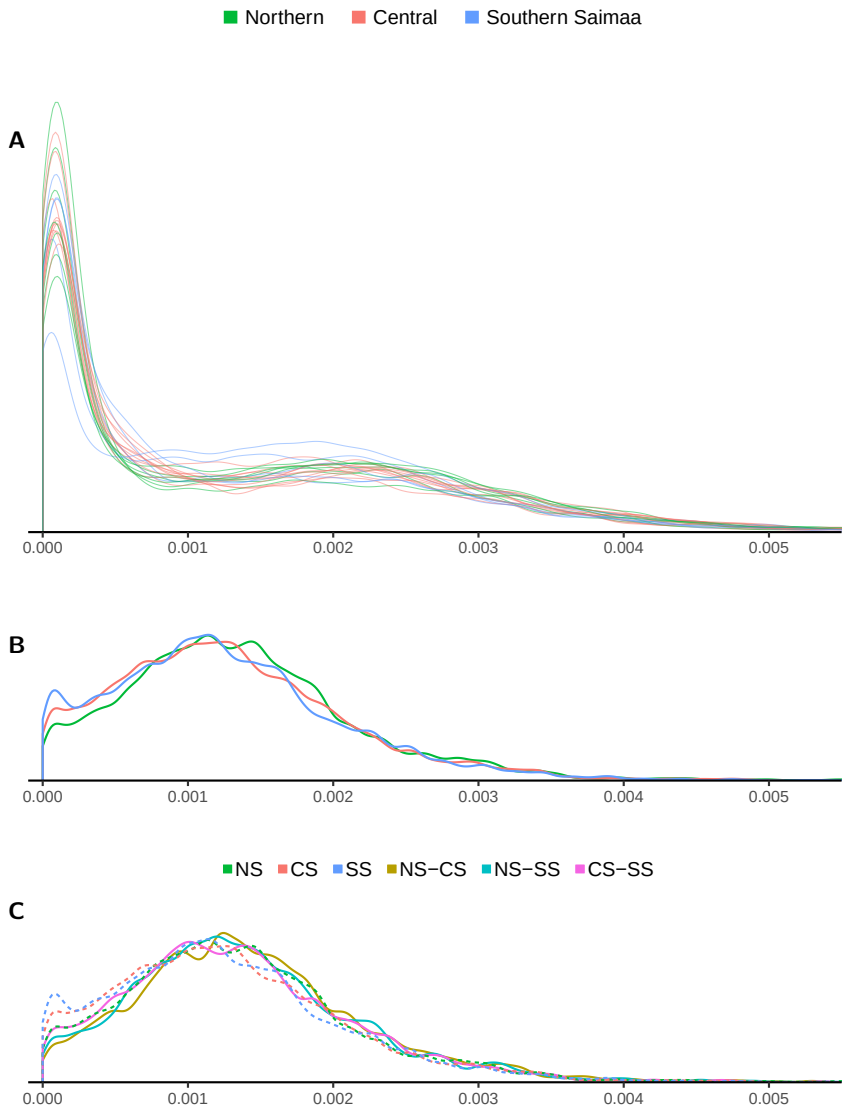

Fig. S10: Nucleotide diversity across the genome. The density lines show nucleotide diversity (A) for the individuals (heterozygosity; H) and (B) for populations (π) in 250 kbp blocks; different parts of Saimaa are indicated with colors. The shift in the distribution of π in comparison to H indicates that the loss of variation has hit different regions in different individuals. (C) The density lines show the mean nucleotide diversity (π; dashed) and the mean nucleotide divergence (d_{xy}; solid) in 250 kbp blocks for different subpopulations (colors). The shift in the distribution of d_{xy} in comparison to π indicates that the loss of variation has hit different regions in different subpopulations.

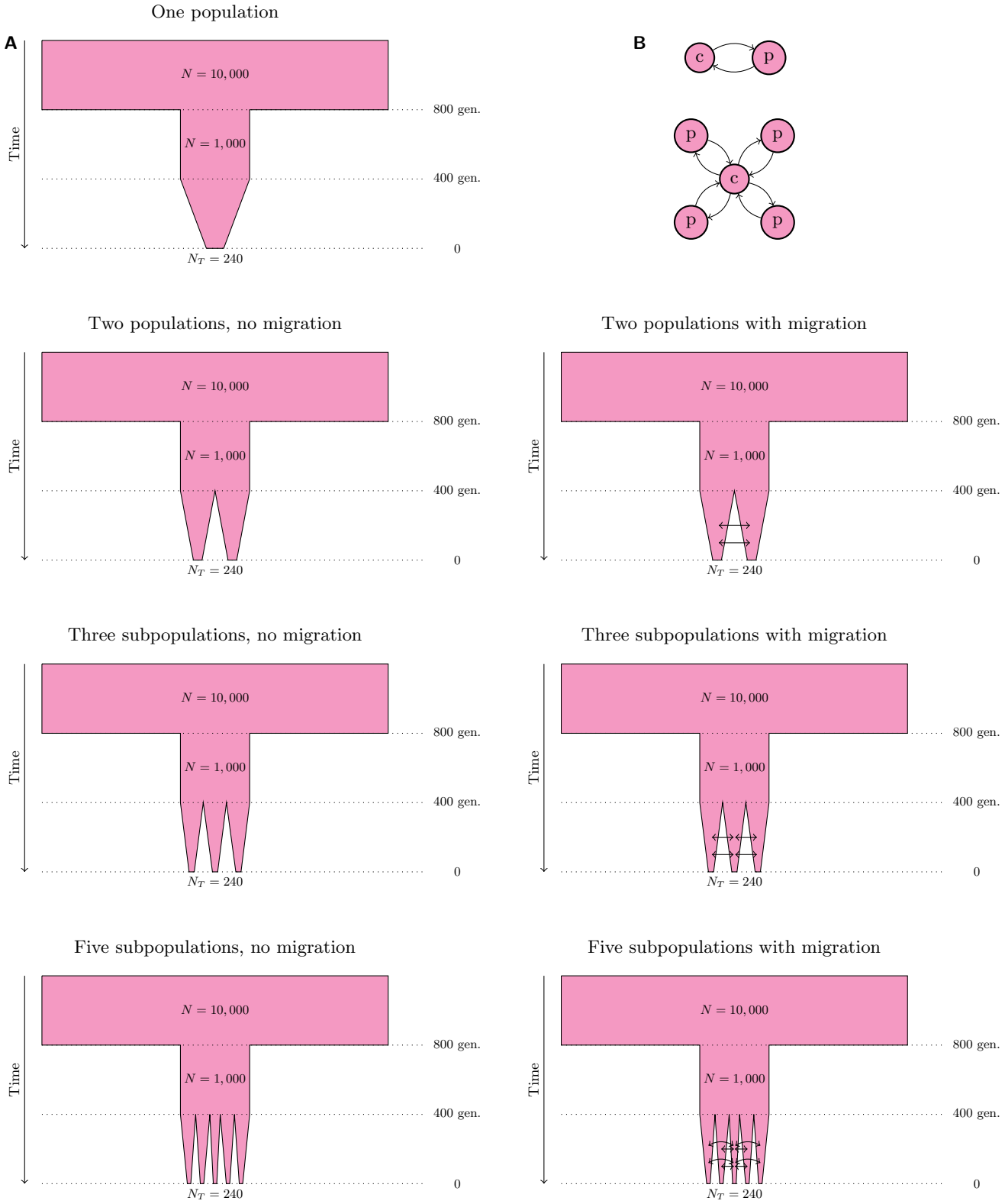

Fig. S11: Simulation models. (A) Some of the demographic models used to simulate synthetic data with msprime. "One population" is a special case as it only models the population size decrease without structuring. Models on the left do not include gene flow while those on the right do. (B) Models with migration have a star-shaped pattern of gene flow between the central (c) subpopulation and varying number of peripheric (p) subpopulations. The one with two subpopulations is a special case as the gene flow is symmetric. Models with two and five subpopulations are shown.

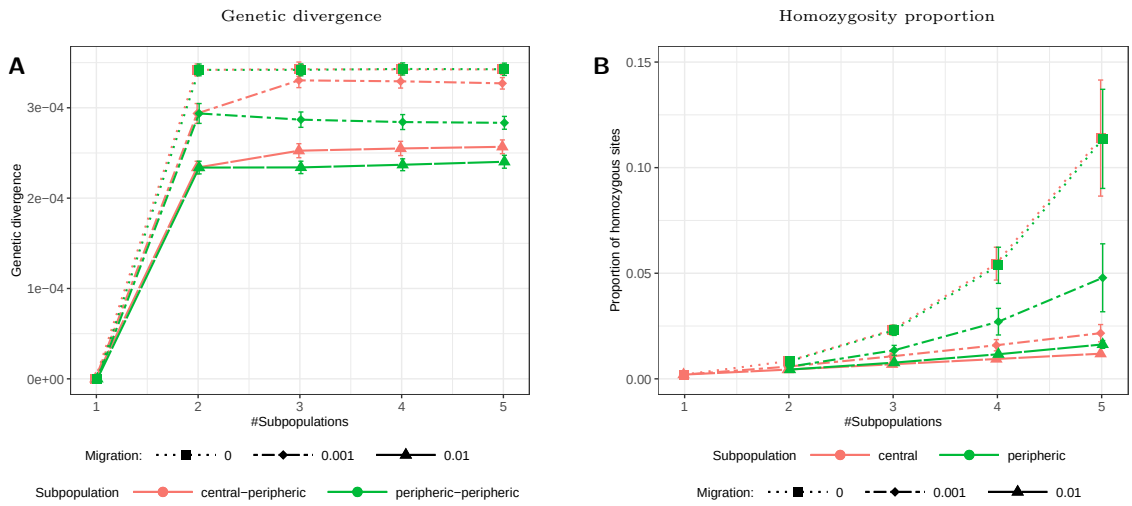

Fig. S12: Preservation of genetic variation in substructured populations. (A) Genetic divergence between subpopulations (y axis) when a population of 240 individuals is either kept as one population (#Subpopulations = 1) or divided into 2-5 subpopulations (x axis) and the migration rate varies (different line types); migration rate of 0 represents fully separated populations. The red and green lines indicate the divergence between the central and the peripheric populations and between different peripheric populations, respectively. (B) Proportion of within-subpopulation homozygous sites of the total segregating sites under different number of subpopulations. Red and green lines indicate central and peripheric populations.

Table S1: Sequencing coverage of the four-subspecies data (left) and the Saimaa data (right). The coverage was computed for the first contig (20.5 Mbp) from bam files, including all sites, and from the vcf file including the called SNPs.

Sample	Population	bam depth	SNP depth
S4-07	Svalbard	9.543	9.257
S6-07	Svalbard	9.711	9.621
S7-07	Svalbard	8.174	8.163
S8-07	Svalbard	9.954	9.652
S11-07	Svalbard	7.764	7.739
S11-96	Svalbard	9.559	8.947
S15-96	Svalbard	8.946	8.483
S18-96	Svalbard	8.486	7.937
S19-07	Svalbard	9.504	9.137
S20-07	Svalbard	23.753	22.913
BG-01	Baltic	9.497	8.827
BG-02	Baltic	8.727	8.054
BG-03	Baltic	25.012	23.106
BG-09	Baltic	11.576	11.124
BG-10	Baltic	8.583	7.957
BG-12	Baltic	10.462	9.777
H2210	Baltic	19.518	18.497
N3-09	Baltic	9.255	8.601
N7-07	Baltic	8.377	7.470
N160	Baltic	2.993	2.253
N210	Baltic	10.508	10.585
NN3-07	Baltic	11.388	10.388
L1501	Ladoga	10.545	9.172
L1683	Ladoga	12.949	11.148
L1684	Ladoga	4.410	3.564
LG-01	Ladoga	26.863	24.744
LG-02	Ladoga	8.870	8.319
LG-03	Ladoga	8.174	7.615
LG-04	Ladoga	8.630	8.152
LG-05	Ladoga	8.490	7.974
LG-06	Ladoga	8.645	8.132
258	Saimaa	13.771	12.579
292	Saimaa	13.649	12.140
819	Saimaa	11.463	10.812
1090	Saimaa	8.663	7.835
1271	Saimaa	7.451	6.732
1482	Saimaa	13.241	12.015
1983	Saimaa	22.637	19.138
2259	Saimaa	6.670	6.458
2343	Saimaa	8.558	7.258
2357	Saimaa	8.169	7.639
2399	Saimaa	11.892	10.473
2436	Saimaa	8.646	7.582
2449	Saimaa	12.885	12.838
2474	Saimaa	11.430	10.789
2496	Saimaa	8.309	7.430
2528	Saimaa	13.333	11.594
2564	Saimaa	8.831	7.931
2585	Saimaa	10.677	9.750
2614	Saimaa	12.738	11.922

Sample	bam depth
297	3.643
298	6.986
1235	4.652
1262	6.206
1270	6.638
1488	9.300
1552	8.174
1956	3.269
1957	7.442
1977	1.260
2119	2.696
2221	1.209
2232	8.215
2233	7.270
2277	6.831
2327	1.113
2332	2.466
2333	0.434
2337	5.460
2340	7.151
2341	4.643
2342	7.152
2348	5.958
2350	6.187
2352	0.836
2353	2.023
2354	1.210
2356	5.142
2363	4.621
2375	6.257
2376	3.183
2377	0.558
2379	8.209
2380	5.550
2382	7.823
2383	5.900
2392	8.311
2393	7.478
2395	1.123
2396	5.503
2397	8.362
2402	3.998
2404	4.262

Sample	bam depth
2405	1.798
2411	4.682
2421	2.225
2422	5.195
2426	4.626
2427	10.610
2432	7.058
2433	5.591
2440	6.053
2445	4.460
2452	2.797
2454	6.967
2456	2.619
2472	5.434
2477	3.938
2485	1.176
2497	3.276
2499	5.857
2500	4.045
2501	6.541
2504	3.594
2506	8.741
2509	5.502
2511	5.242
2513	6.235
2514	5.089
2527	2.055
2549	4.784
2550	7.326
2551	6.412
2555	1.772
2557	5.416
2563	5.927
2565	8.138
2566	8.144
2567	6.741
2568	6.452
2569	9.112
2570	8.488
2572	5.062
2573	4.459
2574	8.231
2575	0.342

Table S2: Estimates of nucleotide diversity (π) with different SNP ascertainment and using phased & imputed data (phased) or unprocessed data (raw vcf). The use of global MAF (i.e. computed over all samples) has little impact on the phased data and estimates with population MAF (applied for each population separately) are nearly identical. However, the ascertainment has a major impact on the raw vcf data and changes the relative order of the populations. The estimates are based on the 364 contigs that are longer 2 Mbp in length, covering 73.6% of the assembly.

	Global MAF		Population MAF	
Population	phased	raw vcf	phased	raw vcf
Svalbard	0.002083	0.002110	0.002084	0.002555
Baltic	0.002148	0.002191	0.002149	0.002456
Ladoga	0.002096	0.002123	0.002097	0.002308
Saimaa	0.001324	0.001173	0.001325	0.001189

Table S3: For each sample, sampling region and year, age at death, and heterozygosity across the longest 100 contigs.

Sample	Region	Year	Age	*H*
819	NS	1985	0	0.001207
1090	CS	1986	0	0.001541
1235	SS	1987	0	0.001337
1262	NS	1988	0	0.001373
1482	NS	1994	0	0.001322
1488	SS	1994	21	0.001412
1552	CS	1995	0	0.001207
1956	NS	1997	0	0.001383
1957	SS	1997	0	0.001084
1977	CS	1997	0	0.001341
1983	NS	1997	0	0.001047
2119	CS	1998	6	0.001070
2221	NS	2000	28	0.001379
2232	NS	2000	5	0.001375
2233	CS	2000	0	0.001578
2259	SS	2001	0	0.001429
2277	NS	2001	NA	0.001508
2327	SS	2002	0	0.001241
2332	SS	2003	1	0.001563
2333	NS	2003	NA	0.001378
2337	NS	2003	12	0.001373
2340	NS	2003	0	0.001234
2341	NS	2003	0	0.001175
2342	NS	2003	0	0.001227
2343	CS	2003	0	0.001352
2348	NS	2003	1	0.001378
2350	NS	2003	0	0.001393
2352	NS	2003	0	0.001435
2353	SS	2004	1	0.001507
2354	SS	2004	1	0.001529
2356	NS	2004	0	0.001360
2357	SS	2004	1	0.001083
2363	NS	2004	0	0.001234
2375	CS	2005	0	0.001302
2376	CS	2005	0	0.001351
2377	NS	2005	0	0.001433
2379	NS	2005	0	0.001526
2380	NS	2005	0	0.000997
2382	NS	2005	0	0.001415
2383	CS	2005	NA	0.001410
2392	NS	2005	0	0.001301
2393	NS	2005	0	0.001483
2395	CS	2006	1	0.001477
2396	NS	2006	5	0.001114
2397	CS	2006	0	0.001210
2399	NS	2006	0	0.001128
2402	CS	2006	0	0.001023
2404	CS	2006	0	0.001289
2405	CS	2006	0	0.001276

Sample	Region	Year	Age	*H*
2411	SS	2006	0	0.001323
2421	CS	2006	16	0.001376
2422	NS	2007	0	0.001338
2426	NS	2007	0	0.001319
2427	CS	2007	0	0.001365
2432	NS	2007	0	0.001298
2433	NS	2007	0	0.001456
2436	SS	2007	0	0.001139
2440	NS	2007	0	0.001350
2445	CS	2007	0	0.001364
2449	SS	2007	6	0.001084
2452	NS	2008	0	0.001134
2454	NS	2008	0	0.001297
2456	CS	2008	0	0.001330
2472	CS	2008	0	0.001192
2474	NS	2009	2	0.001480
2477	CS	2009	0	0.001256
2485	NS	2009	5	0.001438
2496	CS	2009	0	0.001142
2497	CS	2010	3	0.001305
2499	NS	2010	0	0.001119
2500	CS	2010	0	0.001309
2501	NS	2010	6	0.001417
2504	NS	2010	33	0.001305
2506	NS	2010	0	0.001424
2509	CS	2010	0	0.001311
2511	NS	2010	0	0.001244
2513	CS	2010	0	0.001191
2514	CS	2011	20	0.001244
2527	CS	2011	0	0.001201
2528	CS	2011	2	0.001331
2549	SS	2012	0	0.001081
2550	CS	2012	0	0.001394
2551	NS	2012	0	0.001073
2555	SS	2012	0	0.001276
2557	NS	2012	0	0.001406
2563	CS	2013	NA	0.001046
2564	CS	2013	NA	0.001403
2565	CS	2013	NA	0.001232
2566	SS	2013	0	0.001305
2567	CS	2013	0	0.001345
2568	NS	2013	0	0.001479
2569	CS	2013	0	**0.001667**
2570	CS	2013	0	0.001357
2572	CS	2013	0	0.001226
2573	SS	2013	0	0.001193
2574	CS	2013	0	0.001375
2575	SS	2013	0	0.001536
2585	NS	2013	0	0.001432

Table S4: Impact of imputation strategy. The tables show nucleotide diversity within populations (π), heterozygosity of individuals (H), population differentiation (F_{ST}; below diagonal) and genetic divergence between populations (d_{xy}; above diagonal) across the whole genome for the three Saimaa ringed seal subpopulations. The table on the left is computed from data imputed together and the table on the right from data imputed in three batches, the three subpopulations handled separately.

Imputed together			F_{ST} or d_{xy}		
	π	H	North	Central	South
North	0.00124	0.00122		0.00125	0.00125
Central	0.00116	0.00119	0.1140		0.00120
South	0.00116	0.00120	0.0893	0.0862	

Imputed separately			F_{ST} or d_{xy}		
	π	H	North	Central	South
North	0.00124	0.00122		0.00131	0.00127
Central	0.00116	0.00118	0.1188		0.00123
South	0.00113	0.00116	0.0971	0.0931	