# Complex population history of admixed stickleback lineages

Xueyun Feng,*[,1] Juha Merilä,[1,2] and Ari Löytynoja*[,3]
[1]Organismal and Evolutionary Biology Research Programme, Faculty of Biological and Environmental Sciences, University of Helsinki, Helsinki, Finland
[2]Research Division of Ecology and Biodiversity, School of Biological Sciences, The University of Hong Kong, Hong Kong SAR
[3]Institute of Biotechnology, HiLIFE, University of Helsinki, Helsinki, Finland
***Corresponding author:** E-mail: xue-yun.feng@helsinki.fi, ari.loytynoja@helsinki.fi.
**Associate Editor:** N.N.

### Abstract

Introgressive hybridization is an important process contributing to adaptation but challenging to identify, undermining the efforts to understand its role in evolution. On the other hand, many analytical methods assume direct descent from a single common ancestor, and multiple waves of colonisation and admixture can violate their assumptions and lead to seriously biased results. A detailed analysis of 888 whole-genome-sequenced nine-spined sticklebacks revealed a complex pattern of population ancestry involving multiple waves of colonization and introgression across northern Europe. The secondary contact zone between the two recognized lineages was found to be wider than anticipated, displaying a smooth gradient of foreign ancestry with some curious exceptions attributable to freshwater isolates preserving footprints of prehistoric colonisation events. The results provide a back-drop against which introgression and its role in local adaptation can be studied and highlight methodological challenges associated with hidden admixture events.

Key words: Admixture, Introgression, Phylogeography, Sticklebacks.

Article

## INTRODUCTION

Population genomic analyses, such as inferences of selection or local adaptation, typically assume populations descending from a single shared ancestor. However, admixture of populations is common (Moran *et al.*, 2021; Rius and Darling, 2014), and multiple waves of colonisation and hybridization among divergent lineages can produce complex population histories (Flegontov *et al.*, 2019). Moreover, introgressive hybridization is recognized to contribute to adaptation and evolutionary novelties (Hedrick, 2013; Lamichhaney *et al.*, 2018; Marques *et al.*, 2019; Oziolor *et al.*, 2019; Suarez-Gonzalez *et al.*, 2018). The genetic signals of different evolutionary processes are hard to disentangle and, more seriously, unaccounted signals from complex admixture history may seriously mislead analytical methods designed for unadmixed

data (Lawson *et al.*, 2018; Scerri *et al.*, 2018). It is yet unclear how ancestral admixture events affect for instance the estimation of admixture proportions and divergence times among contemporary populations, potentially leading to misinterpretations regarding the reconstructed population histories (but see (Flegontov *et al.*, 2019).

Stickleback fishes are popular model systems to study the genetic underpinnings of evolutionary changes in the wild (Reid *et al.*, 2021). The well-developed genetic resources (Jones *et al.*, 2012) have made the three-spined stickleback (*Gasterosteus aculeatus*) a central model in evolutionary biology and enabled study of biological mechanisms across ecological, population genetic, developmental and molecular levels (reviewed in Reid *et al.* 2021). The related nine-spined stickleback (*Pungitius pungitius*) has also gained foothold as a model to study the genetic basis of local adaptations (e.g. Herczeg *et al.*, 2010; Karhunen *et al.*, 2014; Kemppainen *et al.*, 2021; Wang *et al.*, 2020). Unlike its sister species, the nine-spined sticklebacks commonly populate even the smallest freshwater habitats and are found in tiny ponds showing both parallel adaptations and extreme genetic differentiation (Fang *et al.*, 2021; Guo *et al.*, 2019; Shikano *et al.*, 2010a). Together with the resources developed for other stickleback species, the nine-spined sticklebacks' ability to thrive in extreme habitats makes it an ideal model system for understanding the resilience of small populations under anthropogenic change. However, the difficulty of separating signals of demographic and selective events from each other with genetic data (Fang *et al.*, 2021; Guo *et al.*, 2019; Savolainen *et al.*, 2013; Shikano *et al.*, 2010a, b), in particular when populations are subject to strong genetic drift (Kemppainen *et al.*, 2021; Ovaskainen *et al.*, 2011), poses a challenge to study local adaptation in species living in isolates subject to complex population histories. That a detailed understanding of species' history can facilitate progress towards understanding the origin of adaptive variants is provided by studies of our own species. Indeed, humans are a prime example of local adaptation readily reutilizing existing genetic variation and many of our adaptive features have spread through secondary contact and admixture between diverged populations (Hellenthal *et al.*, 2014) or via introgression from sister species (reviewed in Racimo *et al.* 2015).

The nine-spined sticklebacks' colonization of Northwestern Europe and their population structure has been studied extensively with microsatellites, mtDNA and single nucleotide polymorphisms (SNPs) derived with RAD-sequencing approaches (Aldenhoven *et al.*, 2010; Bruneaux *et al.*, 2013; Guo *et al.*, 2019; Shikano *et al.*, 2010a; Teacher *et al.*, 2011). These studies have revealed two distinct evolutionary lineages, Western European lineage (WL) and Eastern

European lineage (EL), and their potential admixture around the Danish straits, connecting the Baltic Sea through the Kattegat and Skagerrak to the North Sea and the Atlantic, and in the southern parts of the Baltic Sea (Shikano *et al.*, 2010a; Teacher *et al.*, 2011). However, the details of the colonisation history especially in regard to admixture and its extent have not been worked out since earlier. Here, we tackled the question in unprecedented detail utilizing a high-quality reference genome (Varadharajan *et al.*, 2019), and calling SNPs for 888 whole-genome sequenced samples from 45 populations, covering a major part of the species' circumpolar distribution range.

The primary aim of this study was to resolve the detailed phylogeographic history of northern European nine-spined sticklebacks using the latest population genetic methods as applied to whole genome resequencing data. Given the likely secondary contact (Shikano *et al.*, 2010a; Teacher *et al.*, 2011), we also wanted to understand the extent and variability in admixture proportions among the populations across the contact zone. Since preliminary analyses revealed an unexpected population structuring, this motivated us to assess how the admixture and incorrect information about population history affects typical analyses of introgressive hybridization and population differentiation. Our results suggest that thorough understanding of the species' past is required for distinguishing genetic signals of adaptation from demographic ones. In particular, we demonstrate that validation and quantification of population admixture allows studying the evolutionary processes (e.g. barriers to gene flow and sex chromosome evolution) within European nine-spined sticklebacks in unforeseen detail and, crucially, probing the role of introgression in their local adaptation.

## RESULTS

Altogether 888 individuals, 8-31 from each of the 45 populations (Fig. 1a, Table S1), were sampled. The samples were whole-genome sequenced to 10-20X coverage, and variants were called using the standard pipeline consisting of bwa (Li, 2013), samtools (Li *et al.*, 2009) and GATK (McKenna *et al.*, 2010, see Methods). In total 15,217,577 SNPs across the 449 Mbp genome were included in the analyses after quality control (Table S2).

### Nuclear DNA indicates rampant admixture of ancestral lineages

The phylogenetic analysis of whole mtDNA data confirmed the earlier findings (Shikano *et al.*, 2010a; Teacher *et al.*, 2011) and recovered a deep split between the WL and EL (Fig. 1b, Fig. S1). The Finnish and Swedish freshwater samples, the northern Baltic Sea samples and the White Sea samples all had the EL mtDNA type, whereas the Atlantic and Western European samples had the WL type (Fig. 1b). The EL type was found at low frequency in the Kattegat, and the WL type on the German Baltic Sea

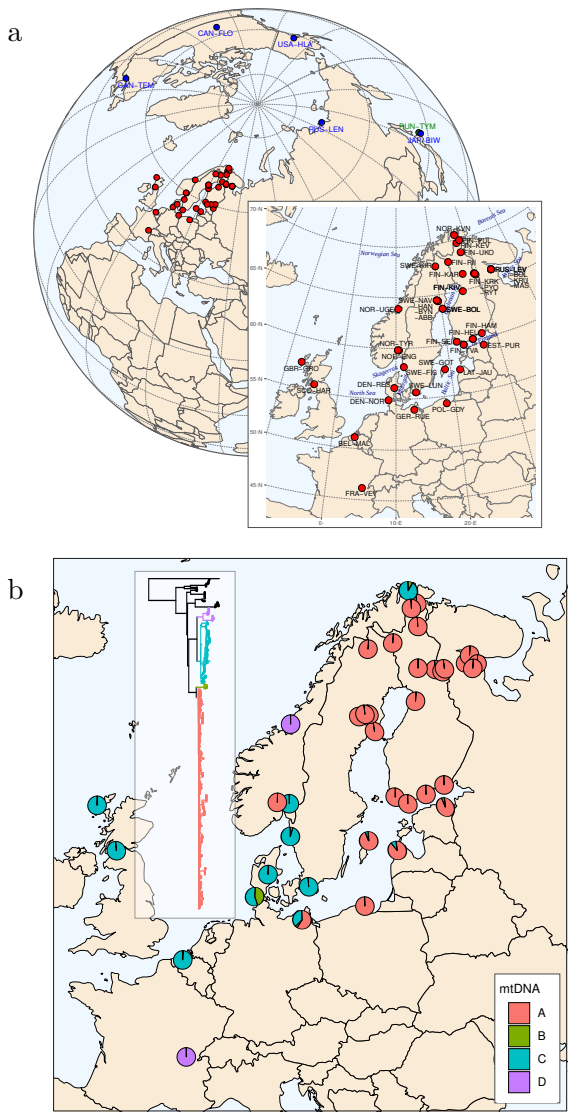


**FIG. 1.** Sampling locations and European mtDNA lineages. a) 8-31 individuals per location were sampled, totalling 888 individuals. Red-dots indicate the focal European populations, blue dots reference populations from other parts of the word. Green dot indicates the origin of the *P. tymensis* out-group individuals. b) The pie charts show the frequency of the four mtDNA types, indicated with distinct colors in the inset tree, for each population. The phylogenetic tree was constructed using RAxML on full mtDNA and rooted with *P. tymensis*.

coast and at low frequency among samples from Latvia and Gotland (Fig. 1b). In addition to clean EL and WL groups, the mtDNA phylogeny contained distinct early branchings at the root of both clades (Fig. 1b, Fig. S1). The mtDNA type branching from the WL clade was found with 100% frequency in isolated French and central Norwegian freshwater populations while the type branching from the EL clade was found in the North Sea, in two Baltic Sea populations and at the Barents Sea coast (Fig. 1b). A clear discrepancy in the distribution of the main EL and WL types was Tyrifjorden (NOR-TYR), a lake population outside Oslo, Norway, that was of EL type with 100% frequency (Fig. 1b).

A nuclear phylogenetic tree was inferred from 1,792 regional Maximum Likelihood trees using ASTRAL (Fig. 2a; Mirarab *et al.* 2014). Consistent with the mtDNA analysis, the nuclear tree grouped Fennoscandian and White Sea populations to EL (blue rectangles) and Atlantic and Western European populations to WL (red dots; Fig. 2a,b). However, in contrast to mtDNA, the nuclear data placed the Baltic Sea (BS; green triangles) populations and several populations from the neighboring Skagerrak/Kattegat (orange triangles) basal to the WL clade (Fig. 2a, Fig. S2). Moreover, these populations formed very short branches as would be expected if they originate from admixture between EL and WL clades (Kopelman *et al.*, 2013; Rheindt and Edwards, 2011). The possibility of a much wider admixture of the two major lineages was supported by population clustering with ADMIXTURE (Alexander *et al.*, 2009) and Principal Component Analysis (PCA). In the ADMIXTURE plot, the intermediate populations showed a mixture of ancestral components associated with the EL and WL populations (Fig. 2c), and the PCA

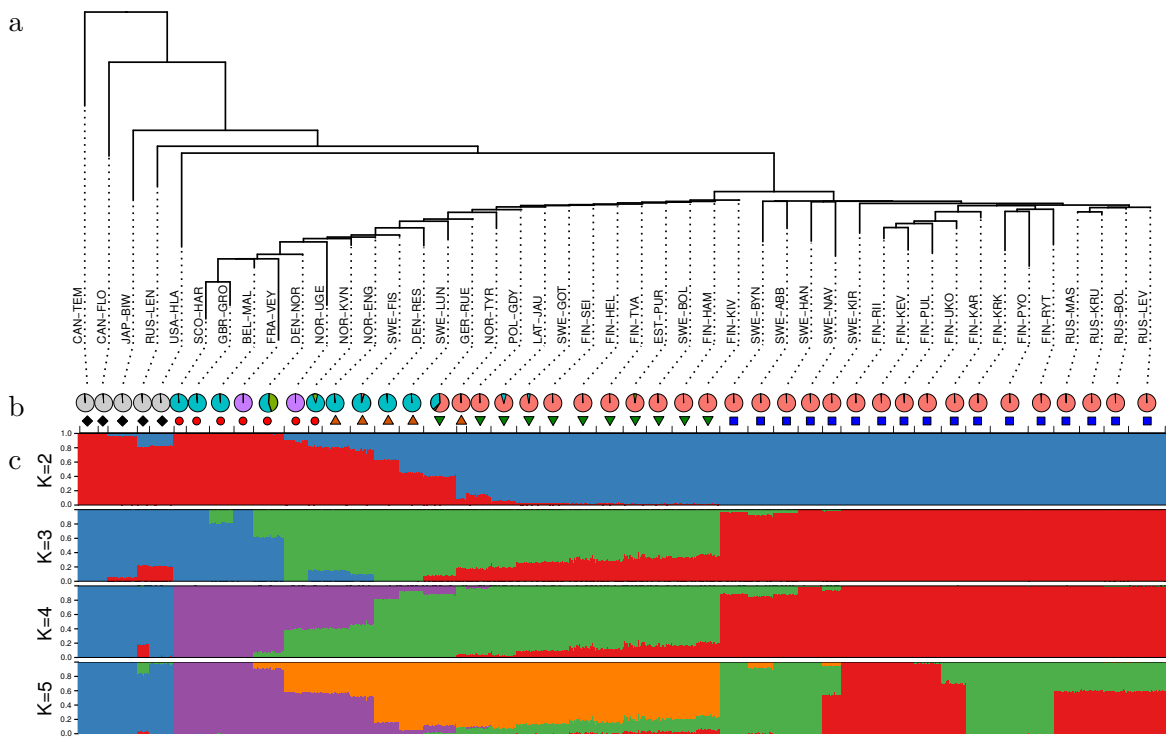


**FIG. 2.** Genetic structure of European nine-spined stickleback populations. a) The nuclear phylogenetic tree was generated from 1,792 regional Maximum Likelihood trees using ASTRAL and rooted with *P. tymensis*. Two individuals were randomly selected from each population. b) The pie charts show the frequency of the four mtDNA types (see Fig. 1b) and the shapes indicate geographic origin of the sampled populations. Black diamonds, red circles, orange triangles, green inverted triangles, and blue squares indicate non-European; Atlantic and Norwegian Sea; Skagerrak/Kattegat; Baltic Sea; and Finnish inland, Swedish inland and White Sea populations, respectively. c) The barplots show the proportion of genetic ancestry of the 888 whole-genome sequenced samples (x-axis) derived from K ancestral populations inferred with ADMIXTURE.

placed these populations as a continuum between the EL and WL populations (Fig. S3). A clear exemption from the current-day geographic distribution was Tyrifjorden (NOR-TYR) that, similarly to mtDNA analyses, grouped among the BS populations (Fig. 2, Fig. S2).

### Ancestral lineages have deep and complex history of admixture

A bifurcating tree cannot fully represent the relationships among introgressed populations. To understand the extent of introgression, we performed an *f-branch* analysis with Dsuite (Malinsky *et al.*, 2021) using the ASTRAL tree (Fig. 2a) as our phylogenetic model. In addition to supporting the earlier findings of admixture in Skagerrak/Kattegat and BS populations, the *f-branch* analysis indicated extensive allele sharing among the BS populations, unsurprising for marine populations connected by gene flow (Fig. S4). More unexpectedly, it unveiled a pattern of introgression from the ancestrally related Northeast Asian and North American populations to most WL populations and to BS populations (Fig. S4). Notably, this signal was not present in the two UK populations. While the positioning of the populations in the phylogenetic model may conceal some signatures of admixture, the result suggest that the WL populations have different

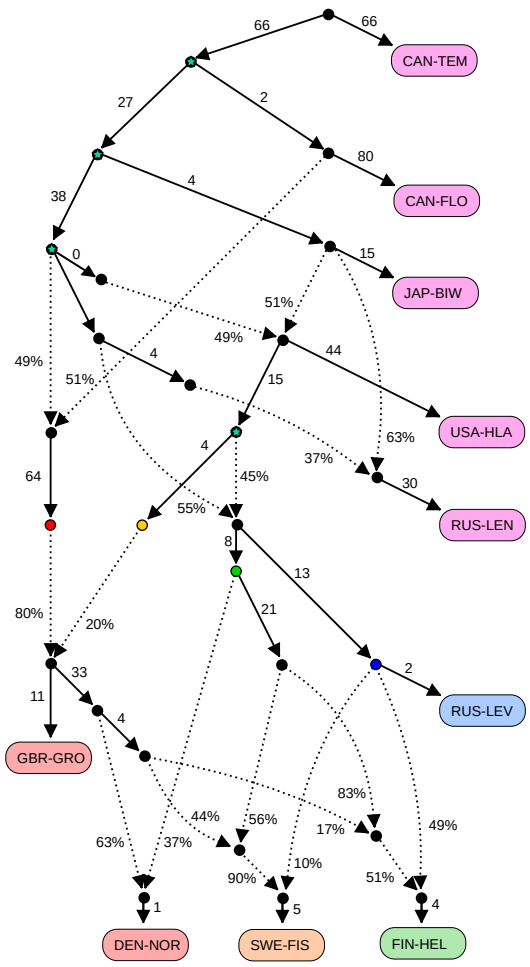


**FIG. 3.** Basal history of European nine-spined stickleback lineages. The optimal qpGraph model (|Z|= 2.7) for the non-European populations (magenta boxes) and the representatives of WL and EL (red and blue, respectively) and two populations intermediate to them (orange and green) shows a complex history of admixture among the lineages forming the current day populations. The UK GBR-GRO descends 80% from a deep ancestor (red dot) but has a 20% contribution from an ancestral lineage (orange dot) related to RUS-LEV from the White Sea. DEN-NOR from the North Sea descends 63% from the old WL but also has a more recent 37% contribution from the east (green dot). RUS-LEV is a mixture of ancestral lineages forming the current day Japanese (JAP-BIW), Alaskan (USA-HLA) and the Lena river (RUS-LEN) populations. SWE-FIS from Skagerrak/Kattegat and FIN-HEL from the Baltic Sea are mixtures of predominantly WL and EL origin, respectively. Numbers next to solid and dotted lines indicate drift (in units of $F_{ST} \times 1{,}000$) and admixture percentages, respectively. Red, orange and green dots indicate hypothetical colonization waves to the Atlantic forming the current-day WL; blue dot indicates the ancestor of the EL that subsequently colonized the northern Fennoscandia. Common ancestors of GBR-GRO and RUS-LEV are marked with cyan stars.

histories and the non-UK populations have been affected by later pulses of migration.

To formally test the latter possibility, we used qpWave (Reich *et al.*, 2012) and qpGraph (Patterson *et al.*, 2012) to infer the history of migration waves. The qpWave analyses indicated at least three and four ancestry streams, respectively, for the different WL and EL populations (see SI Section 1, Tables S3-S6). Given the results, we applied qpGraph to build a model including the early-branching Northeast Asian and North American populations as well as GBR-GRO and RUS-LEV as the representatives of the EL and WL, respectively. According to the best supported model (|Z|=1.6; Fig. S5), the UK population GBR-GRO descends predominantly from two old waves from Asia but has also a minor contribution from a third wave shared with the EL. The White Sea population RUS-LEV descends predominantly from an ancestor that itself was a mixture of lineages forming the current day Japanese and Alaskan populations (Fig. S5). Given this basal model, we added marine populations from the North Sea (DEN-NOR), Skagerrak/Kattegat (SWE-FIS) and the Baltic Sea (FIN-HEL) to the model. According to the best fitting model (|Z|= 2.7; Fig. 3), DEN-NOR is a mixture of the old WL (63%) and a later recolonization lineage (37%). The histories of SWE-FIS and FIN-HEL are more complex: SWE-FIS has less contribution from the old WL than DEN-NOR and has an additional 10% pulse of ancestry from the predecessor of the modern EL; in contrast, FIN-HEL is predominantly of EL origin and has approximately 9% contribution

from the old WL forming the current UK population (Fig. 3).

One should note though that ancestry graphs are sensitive to changes in the input data and often represent only one of many plausible arrangements (Lipson, 2020; Patterson *et al.*, 2012). When we replaced the most recently admixed SWE-FIS and FIN-HEL with an isolated WL population NOR-UGE, the best model ordered the deep branching differently (Fig. S6). With fewer contemporary WL populations in the model, the structure of the graph was simpler, but the different and much deeper ancestry for the UK GBR-GRO was still strongly supported (Fig. S6).

**Secondary contact zone forms a gradient of foreign ancestry**

To explore the population relationships in more detail, we computed the pairwise outgroup-*f3* statistics (Patterson *et al.*, 2012). In this form, the *f3-statistic* measures the shared drift between two test populations in comparison to an outgroup population, the greater value indicating a closer affinity between the test populations. Using a Japanese population (JAP-BIW) as the outgroup, the *f3* analyses revealed major blocks reflecting the Atlantic WL populations, the pure EL populations and a group of intermediate populations from the Baltic Sea and Skagerrak/Kattegat (Fig. 4a). The statistics also showed a finer grain gradient with the more southern BS populations having closer affinity to WL and the more northern BS populations having closer affinity to EL (Fig. 4b,c). Among the WL populations, the two UK populations stood out and showed a greater distance to other populations (Fig. 4a). Repeating the analysis with our most distantly related outgroup population, CAN-TEM from Quebec, Canada, gave qualitatively similar but less resolved results (Fig. S7).

The qpWave, qpAdm and qpGraph analyses indicated that DEN-NOR and RUS-LEV are the optimal WL and EL sources for studying the history of the admixed populations (see SI Sections 1-2, Fig. S8). We used the *f4*-ratio (Patterson *et al.*, 2012; Petr *et al.*, 2019) test to estimate separately the WL and EL ancestry proportions across the different populations (Fig. 5a,b, Tables S8-S9). Using DEN-NOR as the WL source, we saw the WL ancestry decreasing from 68% in the Skagerrak to 63-22% in the Kattegat and the southern Baltic Sea, and then down to 13-11% in the Gulf of Finland and the Bothnian Bay in the north (Fig. 5b). Consistently with this, we saw the EL ancestry decreasing from 69% in the northern Baltic Sea to 45% in the southern parts, and further down to 25-20% in the Skagerrak (Fig. 5b).

In the absence of the earlier analyses of ancestry patterns, one could have acted with caution and selected a WL reference population further from the contact zone. Given the complex structuring among the WL lineage, the use of this alternative reference population would have

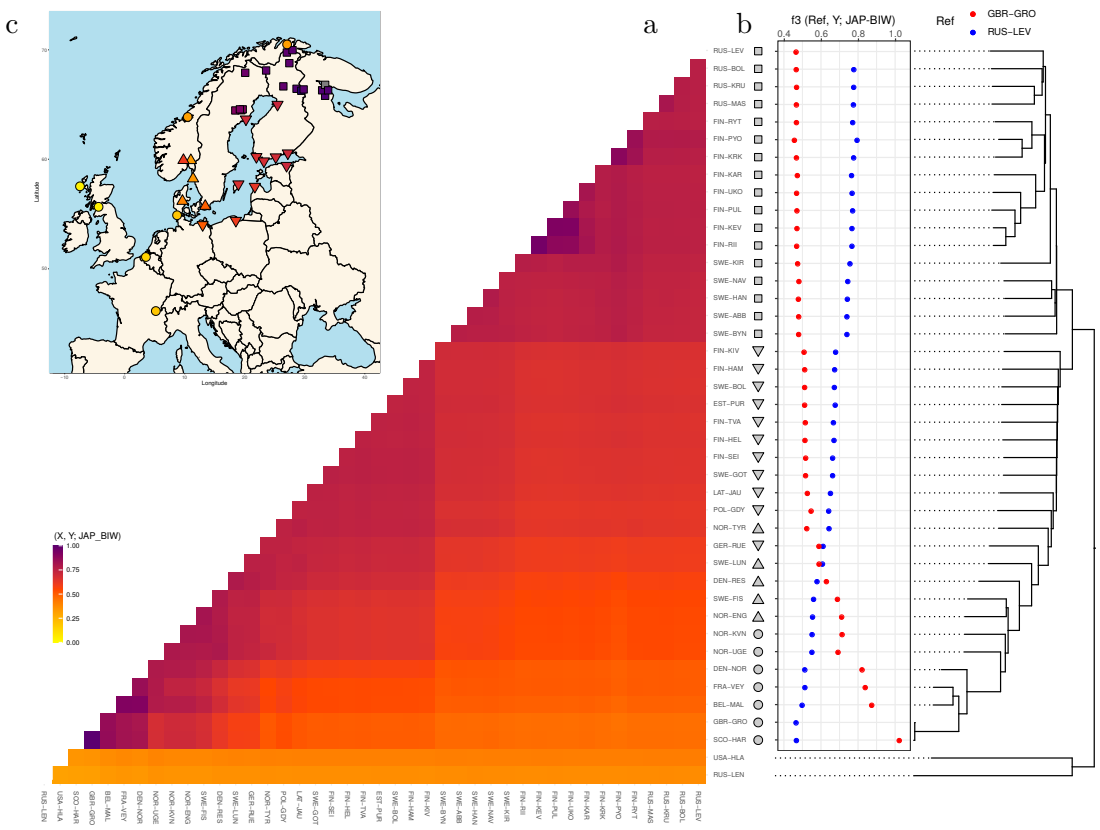


**FIG. 4.** Relationships between WL and EL populations. a) The heatmap colors show the statistics *f3*(X, Y; JAP-BIW), the darker values indicating closer affinity between populations X and Y. The shapes next to the population labels refer to different geographic areas matching those in the insert map. b) The dots indicate the values *f3*(Reference, Y; JAP-BIW) for RUS-LEV (White Sea; red) and GBR-GRO (British Isles; blue). The NJ tree is based on the *f3*-statistics and is rooted with the non-European populations. c) The map shows the geographic location of the populations with shape colors indicating the values *f3*(RUS-LEV, Y; JAP-BIW). RUS-LEV is shown in grey.

indicated that the North Sea population DEN-NOR is introgressed (Table S10). To test how the choice of reference affects the *f4*-ratio test results, we recomputed the statistics using either the UK GBR-GRO or the Belgian BEL-MAL as the WL reference (Tables S8-S9). This indeed had a major impact and most of the WL populations showed significant D-statistic (Table S10) and EL introgression (Fig. 5b, Table S10), naively interpreted to indicate recent gene flow out of the Baltic Sea. Using GBR-GRO as the WL reference, one would conclude that DEN-NOR has 23% of EL ancestry and that the EL gene flow has penetrated deep into central France, FRA-VEY showing 26% of EL ancestry. Estimation of admixture proportions with qpAdm (Lazaridis *et al.*, 2017) gave consistent results (Fig. S8c).

To confirm that such a signal can be created by an unaccounted ancestral admixture within WL, we simulated synthetic data using a population model that reflects the main events inferred to have taken place in the history of European nine-spined sticklebacks (Fig. 5c; see Methods). While the simulation settings were a simplification of the real history, the results nevertheless revealed that the *f4*-ratio analysis is sensitive to unaccounted historical admixture among the

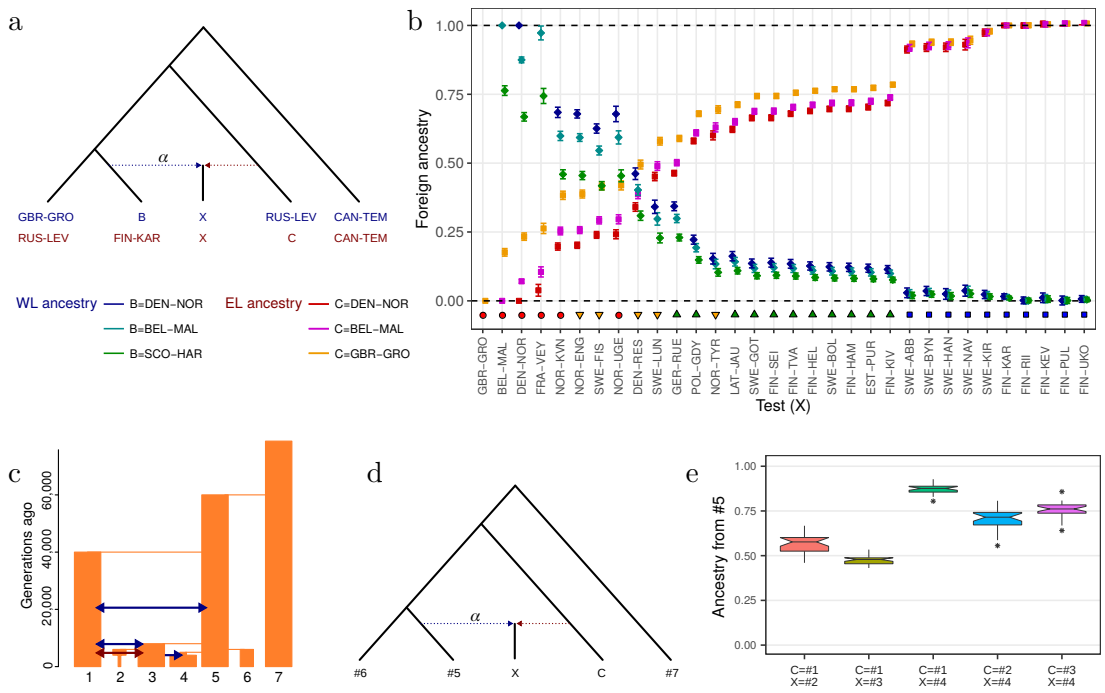


**FIG. 5.** Quantification of WL and EL ancestry in admixed populations. a) The tree depicts the setup of the *f4*-ratio test for assessing the WL ancestry (in blue) and EL ancestry (in red). The WL source (either population B or C) and the test population (X) change between the analyses. The test population is assumed to be a mixture between populations B and C, $\alpha$ indicating the amount of WL or EL ancestry. b) Points show the WL ancestry proportion with 95% confidence intervals for different Test populations and for different WL source population (colors). Ancestry proportions below 0 or over 1 indicate that the test setup is invalid and are not shown. c) The simulation model reflects the inferred main events: population #4 represents an admixed Baltic Sea population founded through a bottleneck from an EL population (#5) and subsequently obtaining 25% of its ancestry from a WL population (#3) that itself had a complex history. Largest populations had $N_e$=100,000, other populations are in scale. Blue and red arrows indicate 25% and 15% of the target population(s), respectively, originating from mass migration. d) The setup of the *f4*-ratio test. e) The correct arrangement, C=#3 and X=#4, gives ancestry proportion very close to the true value of 0.75. The use of population #1 and #2 as a source (C=#1, C=#2) for the test population #4 (X=#4) underestimates and overestimates, respectively, the ancestry from population #5. If #1 is used as a source (C=#1), populations #2 and #3 are inferred to have approximately 50% of ancestry from population #5.

populations (Patterson *et al.*, 2012; Peter, 2016) and the use of an incorrect reference population can inflate or deflate the estimates. The use of the correct reference population, #3 ("DEN-NOR") in our analysis (Fig. 5c,d), gave 1-$\alpha$ estimates very close to the true simulated value of 25% (Fig. 5e). When using the more distantly related population #1 ("GBR-GRO") as the "WL" source in the *f4*-ratio test, we saw also populations #2 ("FRA-VEY") and #3 to obtain high 1-$\alpha$ values (Fig. 5e) despite no recent gene flow from "EL" was simulated (the gene flow from #3 to #4 was unidirectional; Fig. 5c).

**Freshwater isolates provide windows to past admixture events**

Peculiar aberrations in the ancestry analyses were the groupings of NOR-UGE, NOR-TYR, NOR-KVN and NOR-ENG with geographically distant populations (Figs. 4 & 5). To confirm that the observed similarities are real and result from migration after the latest secondary contact rather than from independent admixture events, we computed the rare allele sharing statistics (RASS; Flegontov *et al.*, 2019). This method has been shown to be not affected by strong genetic drift (Flegontov *et al.*, 2019) that often prevails in

population isolates (Savolainen *et al.*, 2013). As the two lineages differ in their sex chromosome systems (Natri *et al.*, 2019), we used the known EL sex chromosome, expected to show a sequencing coverage ratio close to 0.5:1 across the sex-chromosome and pseudo-autosomal parts of LG12 (Fig. S10, Kivikoski *et al.* 2021), to identify the EL males. However, the sex determination system of the WL is not known, and we could not similarly distinguish WL males from WL females.

In line with the *f4*-ratio and the outgroup-*f3* statistics, the two geographically closely located populations from Oslo were clearly differently related to EL and WL, with high levels of rare-allele sharing (RAS) seen between NOR-TYR and BS populations (Fig. S9). Interestingly, NOR-ENG – still connected to Skagerrak – showed higher RAS with NOR-UGE, an isolated WL lake population, than with the nearby marine population SWE-FIS (Fig. S9). The RAS patterns of NOR-KVN and SWE-FIS were nearly indistinguishable, strongly supporting that NOR-KVN is formed by recolonisation from the North Sea/Skagerrak area rather than by an independent local admixture event (Fig. S9). Outside the contact zone, the presence of EL males matched the expected EL and WL ancestries of populations. In the contact zone, exceptions to the expected pattern were the absence of EL males in GER-RUE and the presence of them in DEN-RES, SWE-LUN and NOR-TYR (Fig. 6). While seeing EL sex

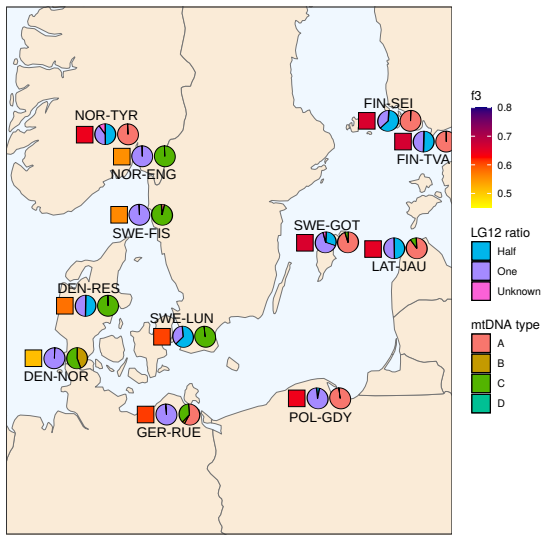


**FIG. 6.** Conflicting genetic patterns in the secondary contact zone. For each population, the colored rectangles show the outgroup-*f3* statistic (Fig. 4b), and the pie charts the LG12 and mtDNA (Fig. 2a) haplotype frequencies. LG12 ratio of "Half" indicates EL male whereas the ratio of "One" can be WL male or WL/EL female.

chromosomes in NOR-TYR was not surprising, DEN-RES and SWE-LUN showed a peculiar combination of EL sex chomosomes, WL mtDNA and predominantly WL nuclear DNA (Fig. 6). Although admixed nuclear DNA and mismatch between the sex chromosomes and mtDNA clearly indicate a mixed ancestry for the two populations, our sampling is too sparse to allow for an accurate reconstruction of the migration and admixture history.

**Admixture can seriously affect the divergence time estimation**

The definition of divergence time is vague when populations have been connected by gene flow multiple times in their history. Nevertheless, we utilized MSMC2 (Malaspinas *et al.*, 2016) and demographic modelling with *moments* (Jouganous *et al.*, 2017) to estimate the divergence times between selected populations and to date the subsequent admixture events. The

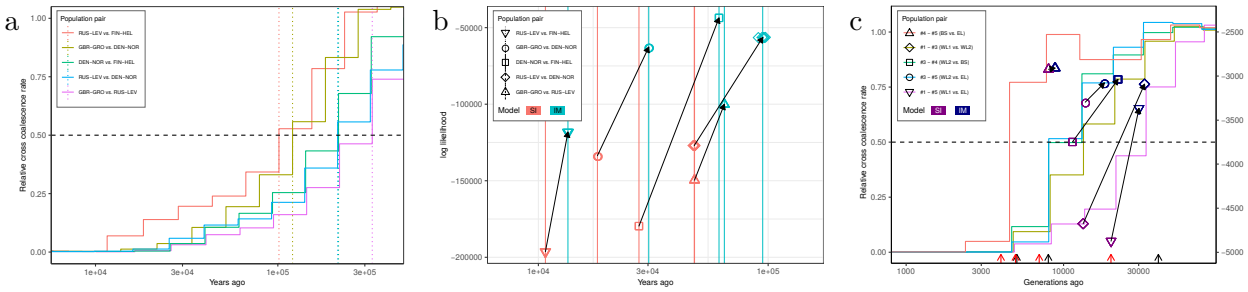


**FIG. 7.** Dating of population separation times. a) Relative cross-coalescence rates for five population pairs representing the different geographic regions. Dotted lines indicate the time at which the relative cross-coalescence rate surpasses 0.5, considered here as the split time for the population pair. b) Log likelihood and separation time for five pairs of populations from *moments* analyses with strict isolation (SI) and isolation with Migration (IM) models with one change in $N_e$. Arrows show the improvement in likelihood and increase in separation time when changing from SI model to IM model. For each pair, five replicates with the best likelihood are shown. Lines and arrow end points represent the mean of those. c) Demographic analyses were replicated with synthetic data simulated using the model in Fig. 5c. Despite the simplistic scenario, the relative cross-coalescence rate results show the gene flow from ancestral population #1 ("WL 1") pushing the divergence time e.g. between populations #3 ("WL 2") and #5 ("EL") back in time. Similar effect is seen in *moments* analyses with strict isolation (SI) and isolation with migration (IM) models, the latter obtaining better likelihoods but in most cases also inferring much deeper divergence times. Black and red arrows at the bottom indicate the times of population divergence and admixture events, respectively.

cross-population analysis with MSMC2 indicated independent history for all studied populations for the last 10,000 years (Fig. 7a). RUS-LEV from the White Sea and FIN-HEL from the Baltic Sea showed the earliest contact, starting around 11,600 years ago, but appeared to have long, partly independent histories, crossing the 0.5 cross-coalescence rate line around 101,400 years ago (Fig. 7a). While all other population pairs showed some contact between 13,800 and 16,400 years ago, DEN-NOR from the North Sea and GBR-GRO from the British Isles united the second fastest, crossing the 0.5 line around 120,400 years ago (Fig. 7a). The pairs containing a WL and an EL population showed much more distant connections and crossed the 0.5 line between 215,200 and 329,500 years ago (Fig. 7a). It is noticeable that the cross-coalescence rate between FIN-HEL and DEN-NOR remains close to zero until 25,700 years ago (Fig. 7a).

For demographic modelling, we built simple two-population models consisting of either presence or absence of migration and assuming constant population size. Unsurprisingly, the models with migration always fitted the data better (Fig. 7b); while improving the likelihood, the models with migration also pushed the split times between the populations further back in time. The change was the smallest between RUS-LEV and FIN-HEL, growing from 10,700 years without migration to 13,500 years with migration, and the greatest between RUS-LEV and DEN-NOR, doubling the split time from 47,700 years to 94,700 years (Fig. 7b). Between the WL and EL populations, the youngest split was estimated for DEN-NOR and FIN-HEL, being 27,400 years ago without migration; however, when migration was included, the split time of 61,100 years ago is not drastically different from that another

of WL-EL pair, GBR-GRO and RUS-LEV, at 64,500 years ago (Fig. 7a).

We applied the divergence time estimation methods also for the synthetic data mimicking the inferred demographic history of the nine-spined sticklebacks. Our *msprime* (Kelleher *et al.*, 2016) model (Fig. 5c; for details, see Script S2) was a simplification of the complex relationships between the populations and, for example, modelled all migration events as one-off pulses (often two-way) of mass migration. While this is clearly unrealistic, it allows setting exact proportions of foreign ancestry. In the MSMC2 analyses, the impacts of gene flow are captured by the most extreme population pairs: The split of populations #4 and #5, mimicking the admixed Baltic Sea population and its parental EL population, is accurately represented by the cross-population coalescence time jumping from 0.1 to 0.75 around 5000 generations ago (Fig. 7c); however, the line does not reach 1, reflecting the deep variation obtained through admixture from the WL. On the other hand, when using 0.5 cross-population rate as the criterion, the deepest simulated divergence between populations #1 and #5, mimicking the earliest WL population and the parental EL population, is underestimated. Moreover, the line deviates from 0 at 6,000 years ago, correctly reflecting the more recent gene flow between the populations (Fig. 7c).

The demographic modelling with *moments* demonstrated the impact of gene flow (Fig. 7c): Strict isolation (SI) gave much younger split times than Isolation with Migration (IM). However, the split time estimated with SI is older than the most recent contact between the populations and the one estimated with IM is younger than the oldest split between the respective lineages. Modelling migration correctly improved the likelihood (Fig. 7c).

## DISCUSSION

An accurate picture of the historical relationships between contemporary populations is the foundation for all evolutionary inference. Due to climatic oscillations, multiple waves of colonisation and admixture are common in the wild and many analytical methods can be seriously misled unless the resulting population structure and demographic history is correctly accounted for (Gompert and Buerkle, 2016; Lange, 2021; Scerri *et al.*, 2018; Theunert and Slatkin, 2017; Vitti *et al.*, 2013). While much of the recent attention in introgression analysis has focused on humans, we show that nine-spined sticklebacks provide an excellent model system to study the mechanisms of admixture, introgression, and the role of deep population structure in shaping present day patterns of genomic differentiation. European nine-spined sticklebacks occur in inter-connected marine populations but also inhabit numerous lakes and ponds formed across Fennoscandia after the last glacier retreat, many of these preserving “genetic snapshots” of the past evolutionary processes.

Unlike bone fragments of archaic humans, stickleback fish allow studying introgression and admixture in action and experimenting with the relicts.

The presence of two diverged evolutionary lineages of European nine-spined sticklebacks is known from previous studies but was thought to be limited to the Danish straits area (Guo *et al.*, 2019; Teacher *et al.*, 2011). Our results confirm the existence and admixture of the two lineages and show that the admixture zone is much wider than previously thought, spreading through the whole Baltic Sea and Skagerrak/Kattegat north of the Danish straits. Basic tools for population structure analyses, such as ADMIXTURE and PCA, were informative about the genetic ancestries of different populations, but failed to disentangle the complex history of migration contributing to the contemporary population structure. Moreover, expansion of the data with additional samples seemed to drastically change the results of our initial D-statistic and *f4*-ratio analyses.

A reinterpretation of the results following outgroup-*f3*, *f-branch*, and qpWave analyses as well as modelling of the population history with qpGraph revealed a much more complex evolutionary history than originally anticipated and massively expanded the scope of our study. The history of the EL populations turned out to be rather simple, supporting recent radiation from an ancestral population related to current day Japanese and Alaskan populations. The colonization of northern Fennoscandia from the White Sea area through a refugia population in the Salla ice lake is well supported by geological evidence (Johansson, 1995). In contrast, we found the WL populations to be deeply structured and derive from multiple colonisation events. Our qpGraph and *f-branch* analyses revealed that the WL and Baltic populations carry ancestries related to the Northeast Asian and North American populations, in line with earlier hypotheses (Guo *et al.*, 2019; Shikano *et al.*, 2010a; Teacher *et al.*, 2011) of multiple waves of migration from Asia towards Europe. This component may originate from a trans-Atlantic event (Aldenhoven *et al.* 2010, but see Guo *et al.* 2019) or from an ancestral migration out of Asia to both Europe and North America (Guo *et al.*, 2019). Our sampling outside Europe was sparse and may have restricted our ability to recover the finer details of the WL colonisation history in Europe. Nevertheless, at least three streams of ancestries, i.e. two old waves from Asia (or North America) and a third, more recent wave shared with the EL, have contributed to contemporary WL. Inclusion of additional non-European samples would probably reveal an even more complex history of migration and admixture.

The population sampling and underlying models are central for genetic parameter estimation, and the inclusion of new WL populations from the UK drastically altered

our estimates of admixture proportions. While the gradient across the Baltic Sea remained, the use of the UK reference population seemed to suggest an "out of Baltic" event and the EL ancestry penetrating deeply into central France. With DEN-NOR, a WL population inferred to be closely related to the true ancestral population by qpGraph, this pattern disappeared. To confirm that the signal is really coming from the population structure within the WL and that the use of DEN-NOR as the WL source of admixture in the Baltic Sea populations is justified, we replicated the analyses with synthetic data reflecting the main events inferred to have taken place in the history of the WL populations. Although the simulation settings were a gross simplification of the real history, the results nevertheless supported the main findings: When the reference populations for the admixture analysis were more distantly related than the true parental populations, the ancestry from the other side (in our case from the EL) was overestimated; similarly to the analyses of real data, the true parental population for the admixture (in our case DEN-NOR) could itself appear admixed.

In addition to ancestry analyses of contemporary populations, the complex history of the WL may also confuse the line between population structure and divergence (Chan *et al.*, 2017; Harrison and Larson, 2014; Supple *et al.*, 2015) and thus complicates phylogeny estimation and affects the divergence time estimates (Hey and Nielsen, 2004; Jones, 2019; Long and Kubatko, 2018; Pinho and Hey, 2010). Classical phylogenetic trees are unsuitable for the representation of histories of admixed populations and our tree on nuclear variation largely reflected the amount of shared ancestries amongst the European populations and showed the typical symptom of admixed populations forming short branches intermediate to the two ancestral lineages (Kopelman *et al.*, 2013; Rheindt and Edwards, 2011). Instead of attempting to date incorrectly inferred tree nodes, we assessed the impact of the population structure on inference of split times with two widely used approaches, the cross-population coalescence rate analysis with MSMC2 (Malaspinas *et al.*, 2016) and the demographic modelling with *moments* (Jouganous *et al.*, 2017).

The results showed that both gene flow and the populations chosen for the analysis as representative of different evolutionary lineages have a major impact on the estimates. With MSMC2 analyses, the earliest contact was seen between RUS-LEV, supposedly a direct descendant of the ancestral EL from the White Sea, and FIN-HEL, an admixed descendant from the Baltic Sea, starting around 11,600 years ago. However, their histories showed partial independence until 100,000 years ago, correctly reflecting the WL contribution to FIN-HEL but indicating potentially an even more complex history than uncovered here. Another notable

result was the history of FIN-HEL and our proxy for its WL parental population, DEN-NOR, that seem to have remained fully isolated until 25,700 years ago (Fig. 7a). If true, this suggests that there have been nine-spined sticklebacks related to the current North Sea populations in the Baltic ice lake (12,600–10,300 BP; Leppäranta and Myrberg 2009), evolving independently from their Atlantic source. If this holds, this would mean the Baltic Sea basin was originally inhabited by WL populations and was only later colonized, and largely taken over, by fish with EL ancestry. This is opposite to our initial thinking that the Baltic Sea is of EL origin and has only rather recently been introgressed by WL gene flow. A similar hypothesis of recolonization of the Baltic Sea from a southern refugium was proposed for the Atlantic salmon (Säisä *et al.*, 2005). Given the complex interactions between the two water bodies (Schwarzer *et al.*, 2008), our estimates may also be affected by the deep ancestry of the WL, pushing the split times further back in time.

The divergence time estimation is highly sensitive to parameters used. We assumed a mutation rate of $1.42\text{x}10^{-8}$ per generation, originally estimated from transcriptome data with a divergence time of 13 Mya between the three- and nine-spined sticklebacks (Guo *et al.*, 2013), and a generation length of 2 years (DeFaveri *et al.*, 2014). Recent studies have pushed the divergence of the two species to 26 Mya (Varadharajan *et al.*, 2019), thus potentially halving the mutation rate and doubling the estimated split times. A detailed determination of the WL-EL split is beyond the scope of this study and the cross-population coalescence analyses and simple demographic models were primarily performed to assess the impact of gene flow on the estimates. Nevertheless, the results underline the difficulty of estimating split times in the case of complex admixture history; in fact, in the case of multiple waves of gene flow, even the definition of split time is unclear. Based on our qpGraph model (Fig. 3), the most recent common ancestor for WL GBR-GRO and EL RUS-LEV was during the latest migration wave from the east and may have inhabited the Arctic Sea, while the earliest split between the lineages possibly happened hundreds of thousands of years earlier at the Pacific.

Our sampling of WL populations was sparse but we could still learn about the colonizations in different time scales. As the WL has received multiple waves of gene flow from the east, we expected NOR-KVN, a population from the coastal Barents Sea in northern Norway to show an intermediate position between the more southern WL populations and the EL. However, NOR-KVN showed close affinity to marine populations from the North Sea and Skagerrak and likely represents a recolonization from the south. On the other hand, isolation to ponds and lakes typically leads to increased drift and loss of genetic variation (Kemppainen *et al.*, 2021; Savolainen *et al.*, 2013; Shikano *et al.*,

2010b). Despite the smaller absolute amounts of variation, the isolates are of interest as they can provide insights into the early evolutionary history by preserving ancestral genetic variation. In our data, NOR-UGE, an isolated lake population from western Norway represents one such relict and is probably a representative of an early state of the latest mixing between the WL and EL lineages; due to its isolation, it has not been affected by the later admixture/migration events and thus provides untainted (but possibly fuzzy) signal not distinguishable in the more connected populations. By definition, the Rare Allele Sharing statistic (RASS) focuses on rare alleles and NOR-UGE shares those with all other WL populations. While lake isolates are not affected by migration, they may be strongly affected by drift and selection, and thus, cannot preserve their ancestral states fully intact. However, adding more relic populations to the analyses in future should provide additional cues about the complex evolutionary history of these fish.

One of the most peculiar findings of our analyses was the drastically different evolutionary history of two populations outside Oslo, Norway. Engervannet (NOR-ENG) is a small pond connected by a 0.5 km long stream to Oslofjord and Skagerrak, and the nine-spined stickleback population there was not genetically different from the nearby marine populations. In contrast, Tyrifjorden (NOR-TYR), an inland lake some 15 km west from Engervannet, had a nine-spined stickleback population most similar to a freshwater population from Latvia (LAT-JAU; Fig. S9). We confirmed this finding with multiple alternative methods, including the outgroup-*f3* and RASS, and are confident that the signal indicates common ancestry, not independent admixture events resulting in similar ancestry proportions. A possible explanation for the Tyrifjorden population's Baltic ancestry is the opening of the Baltic Sea basin at the end of the last ice age. In addition to openings through the Strait of Öresund, there was a corridor across southern Sweden around 11,000 BP (Andrén *et al.*, 2011). While this would have provided a relatively direct path from the central parts of the Baltic Sea to northern parts of Kattegat, the timing may be too early for the ancestral Baltic Sea sticklebacks to be already admixed. However, gene flow from the Baltic Sea to Kattegat is supported by the Kattegat marine populations showing low levels eastern introgression (Fig. 5b, Table S10, Fig. S4). We believe that the 10% contribution from the ancestral EL (blue dot; Fig. 3) to SWE-FIS in our qpGraph analysis also represents such out-of-Baltic gene flow and not yet another migration wave through the north.

While the history of NOR-TYR is enigmatic, other nearby freshwater populations may have even more complex histories. DEN-RES, a population from Silkeborg, Denmark, shares many

rare alleles with the nearby SWE-FIS and SWE-LUN populations, but even more so with NOR-UGE, an isolated pond population from midwest Norway. In RASS, the top five populations for DEN-RES also include LAT-JAU from the Baltic Sea and NOR-KVN from northern Norway. In addition to admixed nuclear DNA, DEN-RES was found to share western mtDNA but EL sex chromosomes. The fact that the two ancestral lineages have different sex chromosome systems (Natri *et al.*, 2019) makes the admixture process and the history of the contact-zone isolates especially intriguing. One should note though that the EL and WL have been shown to produce fertile offspring (Natri *et al.*, 2019), and the divergent sex determination system does not disqualify any of our findings.

The verification of the admixture of two lineages within the Baltic Sea demonstrates that the demographic history indeed has to be carefully considered when comparing populations from different parts of the region. While complicating some analyses, the presence of admixed WL ancestry at low frequencies in some Swedish ponds will provide an excellent opportunity to study the role of introgression in local adaptation. Moreover, the gradient of WL introgression across the Skagerrak/Kattegat and the Baltic Sea provides an exceptional system to study this evolutionary process from different perspectives including roles of genetic incompatibilities and adaptive introgression, in action. Considering that the WL and EL have different sex chromosomes, the nine-spined sticklebacks provide an exceptional system to study the interplay between population admixture and evolving sex chromosomes.

It is also worth pointing out that many non-model systems lack source material (and resources) to study archaic DNA, limiting opportunities to study the ancestral steps of evolutionary processes. Access to the past states through population isolates as in the case of the nine-spined stickleback is a clear asset. Whatever the explanations for the aberrances in populations' genetic and geographic identities are, they hint about the promise the isolated nine-spined stickleback lineages/populations carry: to understand the state of the Baltic Sea in the early stages of the admixture process we can study a snapshot retained in a Norwegian lake; to study the sex chromosome evolution and compatibility of two sex determination systems we can utilize an EL system retained in a Danish river surrounded by populations with the WL system. Our analyses have resolved the big picture of the nine-spined stickleback phylogeography in Northwestern Europe. There are undoubtedly many unanticipated smaller puzzles to study and explore.

## CONCLUSIONS

In conclusion, we have partially untangled a very complex colonisation and admixture history of nine-spined sticklebacks in northern Europe using a suite of advanced population

genomic approaches as applied to whole genome resequencing data. While this is a case study of one particular species, it carries a broader relevance in a number of distinct ways. First, while earlier phylogeographic studies of these populations suggested fairly simple and straightforward scenarios to explain the current patterns genetic affinities among the different populations (Guo *et al.*, 2019; Shikano *et al.*, 2010a; Teacher *et al.*, 2011), our results suggests far more complex history involving repeated admixture events. Since many species have colonized the Atlantic Sea through the Artic Sea (e.g. Laakkonen *et al.*, 2015; Väinölä, 2003; Vermeij, 1991), it seems possible (if not likely) that similar complexities in evolutionary histories underline genetic relationships among populations of many other taxa too. Second, while multiple streams of ancestry in contemporary populations is acknowledged in increasing number of studies (Flegontov *et al.*, 2019; Moreno-Mayar *et al.*, 2018; Posth *et al.*, 2018; Raghavan *et al.*, 2015; Reich *et al.*, 2012; Shirsekar *et al.*, 2021; Skoglund *et al.*, 2015), ancestral populations may also have been admixed or non-panmictic violating the assumptions underlying methods used to infer local adaptation, phylogenetic relationships among populations, as well as their divergence times. Finally, our results reveal the potential incoherence and unpredictability of the population admixture process and demonstrate that the mitochondrial and nuclear DNA may be differently affected in introgression and that the sex chromosomes can be in conflict with the rest of the genome. Crucially, our study resolves the large-scale relationships among the populations and uncovers many finer grain gradients. With the detailed phylogeography of the species, one can start studying the mechanisms and evolutionary factors behind both the expected and unexpected patterns.

## METHODS

### Sampling

The samples used in this study were collected in accordance with the national legislation of the countries concerned. A total of 888 nine-spined stickleback individuals (8–31 per population) were sampled from two earlier identified European evolutionary lineages, the EL (30 populations) and the WL (10 populations), and five ancestral populations from Asia and North America. The species and lineage assignment of populations was based on information from earlier studies (Guo *et al.*, 2019; Teacher *et al.*, 2011), and confirmed with data from this study (see Results). The samples were collected during the local breeding seasons with seine nets and minnow traps. After anesthetizing the fish with an overdose of MS-222, whole fish or fin clips were preserved in 95% ethanol and stored at - 80°C until DNA extraction. In addition, one *P. tymensis*s individual serving as an outgroup was collected from Hokkaido, Japan (43°49'40"N, 145°5'10"E). The sampling sites are shown in Fig. 1; more detailed information,

including sampling site coordinates and dates, sample sizes, population codes, and species names are given in Table S1.

**From sequencing to variant calling**

Extractions of genomic DNA were conducted following the standard phenol-chloroform method (Sambrook and Russell, 2006) from alcohol-preserved fin clips. DNA libraries with an insert size of 300–350 bp were constructed, and 150-bp paired-end reads were generated using an Illumina HiSeq 2500/4000 instrument. Library preparations and sequencing were carried out at the Beijing Genomics Institute (Hong Kong SAR, China) and the DNA Sequencing and Genomics Laboratory, University of Helsinki (Helsinki, Finland).

The reads were mapped to a subset of the nine-spined stickleback reference genome (Varadharajan *et al.*, 2019) using the Burrows-Wheeler Aligner v.0.7.17 (BWA-MEM algorithm, Li 2013) and its default parameters. The subset, called V6b, contained 199 of the original 5,303 contigs and was 449 Mbp long; the removed contigs were inferred to be contamination, haplotypic copies or from the Y chromosome copy of LG12; the included contigs can be found in Data and code availability. Duplicate reads were marked with samtools v.1.7 (Li *et al.*, 2009) and variant calling was performed with the Genome Analysis Toolkit (GATK) v.4.0.1.2 (McKenna *et al.*, 2010) following the GATK Best Practices workflows. In more detail, RealignerTargetCreator and IndelRealigner tools were applied to detect misalignments and realign reads around indels. The command HaplotypeCaller was used to call variants for each individual, parameters set as -stand_emit_conf 3, -stand_call_cof 10, -GQB (10,50), variant index type linear and variant index parameter 128000. The command GenotypeGVCFs was then used to jointly call the variants for all the samples using default parameters. Interspecific variants were removed and binary SNPs were extracted with bcftools v.1.7 (Li *et al.*, 2009), excluding sites located within identified repetitive sequences (Varadharajan *et al.*, 2019). Sites showing low (<8x) average coverage, low (<20) genotype quality and low (<30) quality score were filtered out using vcftools v.0.1.5 (Danecek *et al.*, 2011). For details about subsequent filtering of dataset used in different analyses, see Table S2.

**Analysis of population structure and phylogeny**

Approximate population structure among all study samples was estimated using PCA within the PLINK toolset v.1.90 (Purcell *et al.*, 2007) and ancestry estimation within ADMIXTURE v.1.3.0 (Alexander *et al.*, 2009). In the latter, the analysis was replicated with the number of ancestral populations (K) varying from 2 to 5. In both PCA and ADMIXTURE analysis, variants with pairwise linkage disequilibrium above 0.1 (PLINK; –indep-pairwise 50 10 0.1) were removed.

The mtDNA variant calls were made haploid using the bcftools plugin fixploidy. A maximum

likelihood phylogeny was inferred using RAxML v. 8.2.12 (Stamatakis, 2014), under the GTR-GAMMA model and using the lewis model to account for the ascertainment bias in SNP data (--asc-corr=lewis in RAxML). The branch support values were computed from 1,000 bootstrap replicates. The tree was rooted with *P. tymensis*.

Two samples were randomly selected from each of the 45 *P. pungitius* populations and one *P. tymensis* sample was included as the outgroup. We set the window size as 5,000 SNPs instead of physical length, aiming to have similar strength of phylogenetic signal for each local genomic region. Small windows close to the end of each autosome with less than 5,000 variants were discarded. SNPs were required to present in 50% of the samples with quality filter of minQ30, minDP8, minGQ20; in total, this left 1,792 valid regions to study. The maximum likelihood phylogenies were inferred using RAxML v. 8.2.12 (Stamatakis, 2014) with the same settings as for the mitochondrial data, and the best tree for each region was selected from 20 alternative runs that started from different maximum parsimony trees (the -N 20 option in RAxML). With the 1,792 local best trees, we then reconstructed a species tree under the multispecies coalescent model using ASTRAL v. 5.7.4 (Mirarab *et al.*, 2014) with default settings. The Local Posterior Probabilities (Mirarab *et al.*, 2014) were calculated for the main topology. The tree was rooted with *P. tymensis*.

**Symmetry statistics and admixture tests**

D-statistics, outgroup *f3*-statistics, and *f4*-ratio tests were computed with ADMIXTOOLS v.5.1 (Patterson *et al.*, 2012) using the programs qpDstat, qp3Pop and qpF4ratio with default parameters. We computed the outgroup *f3*-statistics in the form of *f3*(pop1, pop2; outgroup), which measures the shared drift between pop1 and pop2 with higher *f3* values indicating closer affinity between pop1 and pop2 in comparison to the outgroup. The *f-branch* tests were computed using Dsuite v.0.4r38; (Malinsky *et al.*, 2021) with the ASTRAL tree as the input topology. For D-statistics and *f4*-ratio tests, the outgroup was fixed to be the CAN-TEM samples.

The outgroup-*f3* analyses were performed using a Japanese marine population (JAP-BIW) and a deeply diverged Canadian population (CAN-TEM). The two matrices were displayed as heatmap using gtools::heatmap.2 function in R. A Neighbor-joining tree was generated from an inverse distance matrix using the hclust function in R.

Rare variants shared among populations were analysed with RASS (Flegontov *et al.*, 2019). Rare alleles were ascertained based on the five ancestral non-European populations (RUS-LEN, USA-HLA, CAN-FLO, JAP-BIW, CAN-TEM; henceforth: ANC5) and representative populations from WL (GBR-GRO) and EL (RUS-LEV). SNPs with minimum allele count between 2-10 in the reference panel and below 25% missing

data were included. Alleles were polarised using *P. tymensis*.

### Modelling admixture with qpWave, qpAdm and qpGraph

To determine the minimum number of pulses of ancestry contributing to the WL, BS and EL populations, we used the program qpWave (Reich *et al.*, 2012) from ADMIXTOOLS (Patterson *et al.*, 2012). We set the ANC5 populations and selected WL or EL populations as our outgroups, and inferred the minimum number of migration waves to different target populations (see Tables S3 & S6, SI Sections 1-2).

To model the Baltic populations as a mixture of two sources, we first used qpWave to find triplets ($\mathrm{WL}_i$, $\mathrm{EL}_j$, $\mathrm{Baltic}_k$) that were derived from two ancestry streams (Flegontov *et al.*, 2019; Harney *et al.*, 2021) and then used qpAdm (Haak *et al.*, 2015) to infer the optimal source population pair (for details, see Fig. S8, SI Sections 1-2).

To summarise the information from above analyses, we used qpGraph (Patterson *et al.*, 2012) to model the complex relationships between diverse populations. qpGraph assesses the fit of admixture graph models to allele frequency correlation patterns as measured by f-statistics. We started by fitting the deep/ancestral populations and using it as a basal model to gradually add different WL, EL and BS population combinations. Only graph models with no individual *f4*-statistics showing $|Z| \leq 3.5$ between empirical and modelled statistics were retained. For details of each qpGraph model, see SI Section 3.

### Sex chromosome (LG12) sequencing coverage

The ratio of sequencing coverage at the sex-chromosome versus pseudo-autosomal part of LG12 was computed from the variant (vcf) data. For the selection of SNPs, we utilized the reference male (100X coverage) and five known females (each 10X) from the same FIN-PYO population. We first selected SNPs that had 5-15X coverage in each female and 30-75X combined coverage. We pruned down to 220,642 SNPs with coverage <50X in the male as the markers for the sex chromosome part in LG12 (region 1-2e7 bp), and to 430,056 SNPs with coverage <85X in the male as the markers for the pseudo-autosomal chromosome part in LG12 (region 2.75e7-4.1e7 bp). Finally, we computed the ratio of the mean 'INFO/DP' across these regions, and assigned ratios <0.6 as 'Half' and ratios >0.8 as 'One'. The intermediate ratios and samples with less than 250,000 SNPs with coverage >5X in the pseudo-autosomal part were considered as 'Unknown'. Due to atypical sequencing coverage, possibly due to the longer read length (251 nuc vs. 150 nuc in all other samples), the reference male showed the ratio of 0.77 and was assigned as Unknown. On the other hand, the other FIN-PYO samples showed more expected ratios of 0.52-0.53 and 0.94-1.06 for males and females, respectively.

### Estimation of the divergence times and data simulation

The population size histories for selected populations were inferred using MSMC2 v.2.1.1 (Malaspinas *et al.*, 2016) with default parameters and assuming mutation rate of 1.42*10⁻8 per generation (Guo *et al.*, 2013) and generation length of 2 years (DeFaveri *et al.*, 2014). Using these, we then inferred the divergence time for selected population pairs using the cross-population test implemented in MSMC2 (Malaspinas *et al.*, 2016) following (Schiffels and Wang, 2020). Demographic modelling with *moments* (Jouganous *et al.*, 2017) was used to infer the divergence time for the same population pairs while also taking the gene flow into account. Details of data processing, filtering and models used in the two analyses can be found in SI Section 4 "Estimation of divergence times and data simulation", Table S2 and Script S1.

Synthetic data were simulated with *msprime* v.0.7.4 (Kelleher *et al.*, 2016) using a population model that reflects the main events inferred to have taken place in the history of European nine-spined sticklebacks (Fig. 5c; see Script S2). More precisely, we simulated an early WL colonisation followed by a contact with the second-wave populations, mimicking the UK populations in our study (#1 / "GBR-GRO"); and one of the second-wave populations having a long independent history (#2 / "FRA-VEY"), while the other had an additional contact with the first WL population (#3 / "DEN-NOR"). The latter population then provided a pulse of admixture to a bottle-necked EL population, mimicking our Baltic Sea populations (#4 / e.g. "FIN-HEL"), while two other EL populations remained clean (#5, #6 / e.g. "FIN-KAR", "RUS-LEV"). We simulated 25 replicates, each 10 Mbp long, and sampled 20 diploids per population. We converted the variant data to eigenstrat format and analysed them similarly to the empirical data with qpF4ratio. Synthetic data were further used to replicate the analyses with *f4*-ratio, msmc2 and *moments*; for details, see SI Section 4.

## References


Aldenhoven, J. T., Miller, M. A., Corneli, P. S., and Shapiro, M. D. 2010. Phylogeography of ninespine sticklebacks (*Pungitius pungitius*) in North America: Glacial refugia and the origins of adaptive traits. *Molecular Ecology*, 19(18): 4061–4076.

Alexander, D. H., Novembre, J., and Lange, K. 2009. Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, 19(9): 1655–1664.

Andrén, T., Björck, S., Andrén, E., Conley, D., Zillén, L., and Anjar, J. 2011. The development of the Baltic Sea Basin during the last 130 ka. In *The Baltic Sea Basin*, pages 75–97. Springer.

Bruneaux, M., Johnston, S. E., Herczeg, G., Merilä, J., Primmer, C. R., and Vasemägi, A. 2013. Molecular evolutionary and population genomic analysis of the nine-spined stickleback using a modified restriction-site-associated DNA tag approach. *Molecular Ecology*, 22(3): 565–582.

Chan, K. O., Alexander, A. M., Grismer, L. L., Su, Y.-C., Grismer, J. L., Quah, E. S., and Brown, R. M. 2017. Species delimitation with gene flow: A

methodological comparison and population genomics approach to elucidate cryptic species boundaries in Malaysian Torrent Frogs. *Molecular Ecology*, 26(20): 5435–5450.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., *et al.* 2011. The variant call format and VCFtools. *Bioinformatics*, 27(15): 2156–2158.

DeFaveri, J., Shikano, T., and Merilä, J. 2014. Geographic variation in age structure and longevity in the nine-spined stickleback (*Pungitius pungitius*). *PLoS One*, 9(7): e102660.

Delaneau, O., Zagury, J.-F., Robinson, M. R., Marchini, J. L., and Dermitzakis, E. T. 2019. Accurate, scalable and integrative haplotype estimation. *Nature Communications*, 10(1): 5436.

Fang, B., Kemppainen, P., Momigliano, P., and Merilä, J. 2021. Population structure limits parallel evolution in sticklebacks. *Molecular Biology and Evolution*. msab144.

Flegontov, P., Altınışık, N. E., Changmai, P., Rohland, N., Mallick, S., Adamski, N., Bolnick, D. A., Broomandkhoshbacht, N., Candilio, F., Culleton, B. J., *et al.* 2019. Palaeo-Eskimo genetic ancestry and the peopling of Chukotka and North America. *Nature*, 570(7760): 236–240.

Gompert, Z. and Buerkle, C. A. 2016. What, if anything, are hybrids: Enduring truths and challenges associated with population structure and gene flow. *Evolutionary Applications*, 9(7): 909–923.

Guo, B., Chain, F. J., Bornberg-Bauer, E., Leder, E. H., and Merilä, J. 2013. Genomic divergence between nine- and three-spined sticklebacks. *BMC Genomics*, 14(1): 1–11.

Guo, B., Fang, B., Shikano, T., Momigliano, P., Wang, C., Kravchenko, A., and Merilä, J. 2019. A phylogenomic perspective on diversity, hybridization and evolutionary affinities in the stickleback genus Pungitius. *Molecular Ecology*, 28(17): 4046–4064.

Gysels, E., Hellemans, B., Pampoulie, C., and Volckaert, F. 2004. Phylogeography of the common goby, *Pomatoschistus microps*, with particular emphasis on the colonization of the Mediterranean and the North Sea. *Molecular Ecology*, 13(2): 403–417.

Haak, W., Lazaridis, I., Patterson, N., Rohland, N., Mallick, S., Llamas, B., Brandt, G., Nordenfelt, S., Harney, E., Stewardson, K., *et al.* 2015. Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature*, 522(7555): 207–211.

Harney, É., Patterson, N., Reich, D., and Wakeley, J. 2021. Assessing the performance of qpAdm: A statistical tool for studying population admixture. *Genetics*, 217(4): iyaa045.

Harrison, R. G. and Larson, E. L. 2014. Hybridization, introgression, and the nature of species boundaries. *Journal of Heredity*, 105(S1): 795–809.

Hedrick, P. W. 2013. Adaptive introgression in animals: Examples and comparison to new mutation and standing variation as sources of adaptive variation. *Molecular Ecology*, 22(18): 4606–4618.

Hellenthal, G., Busby, G. B., Band, G., Wilson, J. F., Capelli, C., Falush, D., and Myers, S. 2014. A genetic atlas of human admixture history. *Science*, 343(6172): 747–751.

Herczeg, G., Turtiainen, M., and Merilä, J. 2010. Morphological divergence of North-European nine-spined sticklebacks (*Pungitius pungitius*): Signatures of parallel evolution. *Biological Journal of the Linnean Society*, 101(2): 403–416.

Hey, J. and Nielsen, R. 2004. Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*. *Genetics*, 167(2): 747–760.

Johansson, P. 1995. *The deglaciation in the eastern part of the Weichselian ice divide in Finnish Lapland*.

Geologian Tutkimuskeskus.

Jones, F. C., Grabherr, M. G., Chan, Y. F., Russell, P., Mauceli, E., Johnson, J., Swofford, R., Pirun, M., Zody, M. C., White, S., *et al.* 2012. The genomic basis of adaptive evolution in threespine sticklebacks. *Nature*, 484(7392): 55–61.

Jones, G. R. 2019. Divergence estimation in the presence of incomplete lineage sorting and migration. *Systematic Biology*, 68(1): 19–31.

Jouganous, J., Long, W., Ragsdale, A. P., and Gravel, S. 2017. Inferring the joint demographic history of multiple populations: Beyond the diffusion approximation. *Genetics*, 206(3): 1549–1567.

Karhunen, M., Ovaskainen, O., Herczeg, G., and Merilä, J. 2014. Bringing habitat information into statistical tests of local adaptation in quantitative traits: A case study of nine-spined sticklebacks. *Evolution*, 68(2): 559–568.

Kelleher, J., Etheridge, A. M., and McVean, G. 2016. Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLoS Computational Biology*, 12(5): e1004842.

Kemppainen, P., Li, Z., Rastas, P., Löytynoja, A., Fang, B., Yang, J., Guo, B., Shikano, T., and Merilä, J. 2021. Genetic population structure constrains local adaptation in sticklebacks. *Molecular Ecology*, 30(9): 1946–1961.

Kivikoski, M., Rastas, P., Löytynoja, A., and Merilä, J. 2021. Automated improvement of stickleback reference genome assemblies with Lep-Anchor software. *Molecular Ecology Resources*.

Kopelman, N. M., Stone, L., Gascuel, O., and Rosenberg, N. A. 2013. The behavior of admixed populations in neighbor-joining inference of population trees. In *Biocomputing 2013*, pages 273–284. World Scientific.

Korneliussen, T. S., Albrechtsen, A., and Nielsen, R. 2014. ANGSD: Analysis of next generation sequencing data. *BMC Bioinformatics*, 15(1): 356.

Laakkonen, H. M., Strelkov, P., Lajus, D. L., and Väinölä, R. 2015. Introgressive hybridization between the Atlantic and Pacific herrings (*Clupea harengus* and *C. pallasii*) in the north of Europe. *Marine Biology*, 162(1): 39–54.

Lamichhaney, S., Han, F., Webster, M. T., Andersson, L., Grant, B. R., and Grant, P. R. 2018. Rapid hybrid speciation in Darwin's finches. *Science*, 359(6372): 224–228.

Lange, J. D. 2021. *Natural selection in Drosophila melanogaster: A new detection method, impact on demographic inference, and short-term evolution*. Ph.D. thesis, The University of Wisconsin-Madison.

Lawson, D. J., Van Dorp, L., and Falush, D. 2018. A tutorial on how not to over-interpret STRUCTURE and ADMIXTURE bar plots. *Nature Communications*, 9(1): 1–11.

Lazaridis, I., Mittnik, A., Patterson, N., Mallick, S., Rohland, N., Pfrengle, S., Furtwängler, A., Peltzer, A., Posth, C., Vasilakis, A., *et al.* 2017. Genetic origins of the Minoans and Mycenaeans. *Nature*, 548(7666): 214–218.

Leppäranta, M. and Myrberg, K. 2009. *Physical oceanography of the Baltic Sea*. Springer Science & Business Media.

Li, H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv:1303.3997*.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. 2009. The sequence alignment/map format and SAMtools. *Bioinformatics*, 25(16): 2078–2079.

Lipson, M. 2020. Applying f4-statistics and admixture graphs: Theory and examples. *Molecular Ecology Resources*, 20(6): 1658–1667.

Long, C. and Kubatko, L. 2018. The effect of gene flow on coalescent-based species-tree inference. *Systematic Biology*, 67(5): 770–785.

Malaspinas, A.-S., Westaway, M. C., Muller, C., Sousa, V. C., Lao, O., Alves, I., Bergström, A., Athanasiadis, G., Cheng, J. Y., Crawford, J. E., *et al.* 2016. A genomic history of Aboriginal Australia. *Nature*, 538(7624):

207–214.

Malinsky, M., Matschiner, M., and Svardal, H. 2021. Dsuite-fast D-statistics and related admixture evidence from VCF files. *Molecular Ecology Resources*, 21(2): 584–595.

Marques, D. A., Meier, J. I., and Seehausen, O. 2019. A combinatorial view on speciation and adaptive radiation. *Trends in Ecology & Evolution*, 34(6): 531–544.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., *et al.* 2010. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, 20(9): 1297–1303.

Mirarab, S., Reaz, R., Bayzid, M. S., Zimmermann, T., Swenson, M. S., and Warnow, T. 2014. ASTRAL: Genome-scale coalescent-based species tree estimation. *Bioinformatics*, 30(17): i541–i548.

Moran, B. M., Payne, C., Langdon, Q., Powell, D. L., Brandvain, Y., and Schumer, M. 2021. The genomic consequences of hybridization. *Elife*, 10: e69016.

Moreno-Mayar, J. V., Vinner, L., de Barros Damgaard, P., De La Fuente, C., Chan, J., Spence, J. P., Allentoft, M. E., Vimala, T., Racimo, F., Pinotti, T., *et al.* 2018. Early human dispersals within the Americas. *Science*, 362(6419): eaav2621.

Natri, H. M., Merilä, J., and Shikano, T. 2019. The evolution of sex determination associated with a chromosomal inversion. *Nature Communications*, 10(1): 1–13.

Ovaskainen, O., Karhunen, M., Zheng, C., Arias, J. M. C., and Merilä, J. 2011. A new method to uncover signatures of divergent and stabilizing selection in quantitative traits. *Genetics*, 189(2): 621–632.

Oziolor, E. M., Reid, N. M., Yair, S., Lee, K. M., VerPloeg, S. G., Bruns, P. C., Shaw, J. R., Whitehead, A., and Matson, C. W. 2019. Adaptive introgression enables evolutionary rescue from extreme environmental pollution. *Science*, 364(6439): 455–457.

Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., Genschoreck, T., Webster, T., and Reich, D. 2012. Ancient admixture in human history. *Genetics*, 192(3): 1065–1093.

Peter, B. M. 2016. Admixture, population structure, and F-statistics. *Genetics*, 202(4): 1485–1501.

Petr, M., Pääbo, S., Kelso, J., and Vernot, B. 2019. Limits of long-term selection against Neandertal introgression. *Proceedings of the National Academy of Sciences USA*, 116(5): 1639–1644.

Pinho, C. and Hey, J. 2010. Divergence with gene flow: Models and data. *Annual Review of Ecology, Evolution, and Systematics*, 41: 215–230.

Posth, C., Nakatsuka, N., Lazaridis, I., Skoglund, P., Mallick, S., Lamnidis, T. C., Rohland, N., Nägele, K., Adamski, N., Bertolini, E., *et al.* 2018. Reconstructing the deep population history of Central and South America. *Cell*, 175(5): 1185–1197.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., De Bakker, P. I., Daly, M. J., *et al.* 2007. PLINK: A tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 81(3): 559–575.

Racimo, F., Sankararaman, S., Nielsen, R., and Huerta-Sánchez, E. 2015. Evidence for archaic adaptive introgression in humans. *Nature Reviews Genetics*, 16(6): 359–371.

Raghavan, M., Steinrücken, M., Harris, K., Schiffels, S., Rasmussen, S., DeGiorgio, M., Albrechtsen, A., Valdiosera, C., Ávila-Arcos, M. C., Malaspinas, A.-S., *et al.* 2015. Genomic evidence for the Pleistocene and recent population history of Native Americans. *Science*, 349(6250).

Reich, D., Patterson, N., Campbell, D., Tandon, A., Mazieres, S., Ray, N., Parra, M. V., Rojas, W., Duque, C., Mesa, N., *et al.* 2012. Reconstructing native

American population history. *Nature*, 488(7411): 370–374.

Reid, K., Bell, M. A., and Veeramah, K. R. 2021. Threespine stickleback: A model system for evolutionary genomics. *Annual Review of Genomics and Human Genetics*, 22.

Rheindt, F. E. and Edwards, S. V. 2011. Genetic introgression: An integral but neglected component of speciation in birds. *The Auk*, 128(4): 620–632.

Rius, M. and Darling, J. A. 2014. How important is intraspecific genetic admixture to the success of colonising populations? *Trends in Ecology & Evolution*, 29(4): 233–242.

Säisä, M., Koljonen, M.-L., Gross, R., Nilsson, J., Tähtinen, J., Koskiniemi, J., and Vasemägi, A. 2005. Population genetic structure and postglacial colonization of Atlantic salmon (*Salmo salar*) in the Baltic Sea area based on microsatellite DNA variation. *Canadian Journal of Fisheries and Aquatic Sciences*, 62(8): 1887–1904.

Sambrook, J. and Russell, D. W. 2006. Purification of nucleic acids by extraction with phenol: Chloroform. *Cold Spring Harbor Protocols*, 2006(1): pdb–prot4455.

Savolainen, O., Lascoux, M., and Merilä, J. 2013. Ecological genomics of local adaptation. *Nature Reviews Genetics*, 14(11): 807–820.

Scerri, E. M., Thomas, M. G., Manica, A., Gunz, P., Stock, J. T., Stringer, C., Grove, M., Groucutt, H. S., Timmermann, A., Rightmire, G. P., *et al.* 2018. Did our species evolve in subdivided populations across Africa, and why does it matter? *Trends in Ecology & Evolution*, 33(8): 582–594.

Schiffels, S. and Wang, K. 2020. MSMC and MSMC2: The multiple sequentially Markovian coalescent. *Methods in Molecular Biology*, 2090: 147–166.

Schwarzer, K., Ricklefs, K., Bartholomä, A., and Zeiler, M. 2008. Geological development of the North Sea and the Baltic Sea. *Die Küste, 74 ICCE*, (74): 1–17.

Shikano, T., Shimada, Y., Herczeg, G., and Merilä, J. 2010a. History vs. habitat type: Explaining the genetic structure of European nine-spined stickleback (*Pungitius pungitius*) populations. *Molecular Ecology*, 19(6): 1147–1161.

Shikano, T., Ramadevi, J., and Merilä, J. 2010b. Identification of local-and habitat-dependent selection: Scanning functionally important genes in nine-spined sticklebacks (*Pungitius pungitius*). *Molecular Biology and Evolution*, 27(12): 2775–2789.

Shirsekar, G., Devos, J., Latorre, S. M., Blaha, A., Dias, M. Q., Hernando, A. G., Lundberg, D. S., Burbano, H. A., Fenster, C. B., and Weigel, D. 2021. Fine-scale population structure of North American *Arabidopsis thaliana* reveals multiple sources of introduction from across Eurasia. *bioRxiv*.

Skoglund, P., Mallick, S., Bortolini, M. C., Chennagiri, N., Hünemeier, T., Petzl-Erler, M. L., Salzano, F. M., Patterson, N., and Reich, D. 2015. Genetic evidence for two founding populations of the Americas. *Nature*, 525(7567): 104–108.

Stamatakis, A. 2014. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9): 1312–1313.

Suarez-Gonzalez, A., Lexer, C., and Cronk, Q. C. 2018. Adaptive introgression: A plant perspective. *Biology Letters*, 14(3): 20170688.

Supple, M. A., Papa, R., Hines, H. M., McMillan, W. O., and Counterman, B. A. 2015. Divergence with gene flow across a speciation continuum of Heliconius butterflies. *BMC Evolutionary Biology*, 15(1): 1–12.

Teacher, A. G., Shikano, T., Karjalainen, M. E., and Merilä, J. 2011. Phylogeography and genetic structuring of European nine-spined sticklebacks (*Pungitius pungitius*)—mitochondrial DNA evidence. *PLoS One*, 6(5): e19476.

Theunert, C. and Slatkin, M. 2017. Distinguishing recent admixture from ancestral population structure. *Genome Biology and Evolution*, 9(3): 427–437.

Väinölä, R. 2003. Repeated trans-Arctic invasions in littoral bivalves: Molecular zoogeography of the *Macoma balthica* complex. *Marine Biology*, 143(5): 935–946.

Varadharajan, S., Rastas, P., Löytynoja, A., Matschiner, M., Calboli, F. C., Guo, B., Nederbragt, A. J., Jakobsen, K. S., and Merilä, J. 2019. A high-quality assembly of the nine-spined stickleback (*Pungitius pungitius*) genome. *Genome Biology and Evolution*, 11(11): 3291–3308.

Vermeij, G. J. 1991. Anatomy of an invasion: The trans-Arctic interchange. *Paleobiology*, 17(3): 281–307.

Vitti, J. J., Grossman, S. R., and Sabeti, P. C. 2013. Detecting natural selection in genomic data. *Annual Review of Genetics*, 47: 97–120.

Wang, Y., Zhao, Y., Wang, Y., Li, Z., Guo, B., and Merilä, J. 2020. Population transcriptomics reveals weak parallel genetic basis in repeated marine and freshwater divergence in nine-spined sticklebacks. *Molecular Ecology*, 29(9): 1642–1656.

## ACKNOWLEDGEMENTS

We thank Victor Berger, Pär Byström, Lasse Fast Jensen, Jacquelin De Faveri, Gabor Herczeg, Tuomas Leinonen, Scott McCairns, Andrew McColl, Heini Natri, Takahito Shikano and, Joost Raeymaekers for help in providing samples, and Miinastiina Issakainen, Sami Karja, Laura Häkkinen and Kirsi Kähkönen for help in the laboratory; Paolo Momigliano for help in *moments* anslyses. The advice and support from Antoine Fraimout, Baocheng Guo, Bohao Fang, Cui Wang, Paolo Momigliano, Pasi Rastas, Petri Kemppainen, Zitong Li, Mikko Kivikoski, Jarkko Salojärvi, Jack Beresford, Jacquelin De Faveri, Takahito Shikano, Simon Martin, and Martin Petr is gratefully acknowledged. Our research was supported by grants from the Academy Finland (# 129662, 134728 and 218343 to JM; # 322681 to AL), Helsinki Lifesciences Center (HiLife; to JM), China Scholarship Council (# 201608520032 to XF), and Finnish Cultural Foundation (#00210295 to XF). Computational resources provided by the CSC–IT Center for Science, Finland, are acknowledged with gratitude.

## AUTHOR CONTRIBUTIONS

J.M. started the project. X.F., J.M. and A.L. devised the research idea. X.F. performed the analyses with the help of A.L. X.F. and A.L. wrote the first draft and all authors participated in the writing of the final manuscript.

## ADDITIONAL INFORMATION

**Competing interests:** The authors declare no competing interests.

**Data and code availability:** Raw sequence data will be available at European Nucleotide Archive (ENA) (https://www.ebi.ac.uk/ena) under accession code XXX. The list of contigs included in the reference genome version 6b as well as the code and scripts used in the analyses are available at will be available at https://github.com/XueyunF/nsp_phylogeo. Other relevant data are available from the corresponding author upon request.

## Supplementary Information


1 Analysis of ancestry streams for WL, EL and BS populations . . . . . 29
2 Modelling of Baltic populations as mixture between two lineages . . . . . 30
3 qpGraph modelling . . . . . 31
4 Estimation of divergence times and data simulation . . . . . 32


## List of Figures


S1 Maximum likelihood tree on the mtDNA data . . . . . 33
S2 Cytonuclear incongruence . . . . . 34
S3 Principal Component Analysis of nuclear variant data . . . . . 35
S4 *f-branch* analysis of excess of shared derived alleles . . . . . 36
S5 qpGraph model for basal populations . . . . . 37
S6 qpGraph model for selected European populations . . . . . 38
S7 Outgroup-3 analysis with CAN-TEM as outgroup . . . . . 39
S8 Identification of optimal source pairs with qpWave and qpAdm . . . . . 40
S9 Rare Allele Sharing Statistics . . . . . 41
S10 Identification of EL males . . . . . 42


## List of Tables


S1 Sample information, mtDNA type (A-D) and lineage assignment. . . . . 43
S2 Data filtering for different analyses. . . . . 44
S3 Detailed results of qpWave analysis: WL, EL and Baltic Sea populations. . . . . 45
S4 Detailed results of qpWave analysis: WL and EL populations. . . . . 46
S5 Detailed results of qpWave analysis: WL populations. . . . . 47
S6 Two population symmetrical test with qpWave, the outgroups was set as ANC5. . . . . 48
S7 Detailed results qpWave testing of triples for the Baltic sea populations. . . . . 48
S8 Detailed results of *f4*-ratio test with the EL as source. . . . . 48
S9 Detailed results of *f4*-ratio test with the WL as source. . . . . 48
S10 Detailed results of D-statistics. . . . . 48


## List of Scripts


S1 Python-code for demographic analyses with *moments* . . . . . 49
S2 Python-code for data simulation with *msprime* . . . . . 49

## Analysis of ancestry streams for WL, EL and BS populations

We used qpWave (Reich *et al.*, 2012) from ADMIXTOOLS (Patterson *et al.*, 2012) to estimate the minimum number of pulses of ancestry contributing to the WL, BS and EL populations. The following set of populations were defined as outgroups for the analyses: CAN-TEM, CAN-FLO, JAP-BIW, RUS-LEN, USA-HLA; henceforth called ANC5.

Using ANC5 and RUS-LEV, a marine population from the White Sea, as the outgroups, we found the minimum number of migration pulses to be three for the nine WL populations (GBR-GRO, SCO-HAR, FRA-VEY, BEL-MAL, NOR-UGE, DEN-NOR, SWE-FIS, NOR-KVN and NOR-ENG; p_rank2 = 0.209, for details, see Table S3). According to the saturated model, two pulses originated from populations closely related to our Alaskan USA-HLA and central Canadian CAN-FLO, while the last pulse appeared to be from a population related to the current White Sea population RUS-LEV (Fig. 3). Importantly, the last pulse from RUS-LEV probably reflects a migration from a common ancestral population arriving from the east rather than introgression from the current EL populations. Using the same source set, we first tested the two British populations (GBR-GRO and SCO-HAR) and found the minimum number of migration pulses to be two (p_rank1 = 1, Table S4), with one of the likely sources being a population closely related to the Alaskan USA-HLA. When fixing the British populations in the target set and individually including the seven other WL populations, we found that only the populations from Belgium and central France did not show increased rank (p_rank1 = 0.07-0.08, Table S4), indicating that they can be explained with the same two migration pulses. All the other WL populations showed significant change in the rank value (p_rank2 = 1, Table S4 and S5), meaning that these populations have received an additional pulse of migration which was not experienced by the other WL populations.

Using ANC5 and the British GBR-GRO as the outgroups set, we quantified the number of gene pools for the 17 EL populations from the White Sea and the Finnish and Swedish ponds and lakes. We found the minimum number of migration waves to be four (p_rank3 = 0.828, Table S3). To elucidate the complex signature within EL populations, we performed two and three population qpWave tests. Of these, the White Sea (RUS-LEV/KRU/MAS/BOL) and the nearby Finnish pond populations (FIN-PYO/KRK/RYT/UKO/PUL) showed evidence of a single pulse (p_rank0 > 0.05, Table S4) and symmetrical relationship to the outgroup set. Interestingly, the northern Finnish population FIN-KEV showed an extra pulse of gene flow (Table S4). The Swedish pond populations (except for SWE-HAN with three pulses) and the Finnish population FIN-KAR showed two pulses from the ancestral populations. This is consistent with these now isolated freshwater populations having at some stage been connected to the Baltic Sea basin and receiving WL-like ancestry from there.

Using ANC5, the British GBR-GRO and the White Sea RUS-LEV as the outgroups set, we quantified the number of gene pools for the 12 Baltic Sea populations (FIN-KIV, FIN-HAM, SWE-BOL, EST-PUR, FIN-TVA, FIN-HEL, FIN-SEI, SWE-GOT, LAT-JAU, POL-GDY, GER-RUE, SWE-LUN). The minimum number of migration waves were found to be three (p_rank2= 0.071, Table S3). Removing southern Baltic sea populations (GER-RUE, POL-GDY, SWE-LUN, SWE-GOT) and the isolated Latvian population (LAT-JAU) decreased the migration waves to two (p_rank1=0.144, Table S3). This may indicate migration from a refugia population from the south into the Baltic sea (Gysels *et al.*, 2004; Teacher *et al.*, 2011).

### Modelling of Baltic populations as mixture between two lineages

If a triplet can be modeled as derived from two sources of ancestry in qpWave, it follows that one of the populations can be modeled as a mixture of the other two populations in qpAdm. Based on the phylogeny (Fig. 2b) and D-statistic results (Table S10), the White Sea marine population RUS-LEV was chosen to represent the EL. We performed two way admixture models in a group of three populations using qpWave (Reich *et al.*, 2012) and qpAdm (Haak *et al.*, 2015). First, using qpWave we identified triplets ($\mathrm{WL}_i$, $\mathrm{EL}_j$, $\mathrm{Baltic}_k$) that can explain the Baltic Sea (BS) populations as derived from two ancestry streams (i.e. prank1_tail $> 0.05$, Fig. S8a and Table S7). Second, the validity of the triplets and the chosen EL and WL populations as the optimal source pair was assessed based on qpAdm analyses of other admixed populations (Fig. S8b).

Given the results, the following populations were rejected to be optimal WL source population:

- The British Isle (GBR-GRO) and Scottish population (SCO-HAR) were rejected, as they strongly deviated from the rest of the WL populations and could not be used to distinguish the shared EL-like ancestry present in BS and other WL populations. Although they can be informative for study of the ancient genetic exchange between WL and EL, they are not a good proxy to study the recent admixture in the Baltic sea.
- The central France (FRA-VEY) and Belgium (BEL-MAL) population were rejected, because 1) the two populations were found to be symmetrically related to GBR-GRO and SCO-HAR (Tables S4 & S6); and 2) several BS populations were rejected in the qpAdm analysis (Fig. S8b).
- The isolated Norwegian population (NOR-UGE) was rejected due to the rejection of SWE-FIS and NOR-ENG in the qpAdm analysis (Fig. S8b). However, they have clearly been admixed recently.

We cannot exclude the possibility of a better proxy to study the WL ancestry in the BS populations. However, with our sampling, DEN-NOR was the best candidate: 1) given the uniqueness of DEN-RES

and NOR-TYR (see Discussion in the main text), they were expectedly rejected by qpWave and qpAdm; 2) the *f4*-ratio test (Fig. 5) supports DEN-NOR as a reasonable source and correctly models FRA-VEY as un-admixed; and 3) qpGraph model (Fig. 3) supports DEN-NOR as the best proxy.

## qpGraph modelling

We investigated the ancestry of different populations using modelling with qpGraph. In brief, for given allele frequency data and an input graph topology, qpGraph infers branch length and admixture proportions that minimize the differences between the observed and expected *f4*-statistics.

For the basal qpGraph model (Fig. S5), we used ANC5 as well as GBR-GRO (for WL) and RUS-LEV (for EL) to represent the two lineages. The fitted graph indicated that the Canadian lake population (CAN-FLO) is related to the British WL population. This can be a result of an ancient trans-Atlantic migration (Aldenhoven *et al.*, 2010), or derived from an ancestral population that colonised both North America and Europe (Guo *et al.*, 2019). In line with the hypothesis of multiple waves of out-of-Asia migration, the eastern RUS-LEV and Alaskan USA-HLA populations were shown to derive over 50% of their ancestry from a lineage leading to Japanese JAP-BIW, and the rest from an ancestral population that is distantly related to the Japanese population.

Based on the basal model, we added one population from the North Sea (DEN-NOR), the Skagerrak/Kattegat (SWE-FIS) and the Baltic Sea (FIN-HEL). The three populations were chosen based on qpWave and qpAdm results to test if: 1) there was the North Sea – Baltic Sea connection; and 2) DEN-NOR is the optimal WL source population to model the admixed marine populations. Regardless of the order of populations being added, the model shown in Fig. 3 was the only optimal model.

Given the peculiar patterns observed in NOR-UGE population, we set to test if it is indeed a relic population separated early in the admixture process. The optimal model found by adding NOR-UGE and DEN-NOR into the basal qpGraph model is shown in Fig. S6a. In this model, DEN-NOR and NOR-UGE share a common WL-like ancestral population (deriving 62% and 33% of their ancestry, respectively), but NOR-UGE also has received 67% of its ancestry from a EL-like ancestral population whereas this lineage only contributed 38% of the DEN-NOR ancestry.

To explore the relationships among the WL populations, we excluded the ANC5 set and included the White Sea population (RUS-LEV) and the British Isles population (GBR-GRO) as the representatives for the EL and WL, respectively, and added the four Atlantic WL populations into this model. The optimal model (Fig. S6b) shows that DEN-NOR and NOR-UGE received 53% and 31% of their ancestry from a common WL-like ancestral population. A descendant of this lineage then contributed to 61% of

BEL-MAL and 55% of FRA-VEY ancestry. On the other hand, NOR-UGE and FRA-VEY inherit 69% and 45% of their ancestry from a EL-like population; this evolved further and contributed 47% and 39% of ancestry in DEN-NOR and BEL-MAL. The model is congruent with other lines of evidence showing that NOR-UGE and FRA-VEY are of earlier origin than the other (here, DEN-NOR and BEL-MAL) WL populations.

### Estimation of divergence times and data simulation

We used MSMC2 to infer the split time between populations GBR-GRO, DEN-NOR, RUS-LEV and FIN-HEL. Sample specific VCF and mask files were first created of aligned reads (in bam format) using bamCaller.py from msmc-tools (available at https://github.com/stschiff/msmc-tools) and applying the mappability mask of Kivikoski *et al.* 2021. The variants were then phased with SHAPEIT v.4.2 (Delaneau *et al.*, 2019) and the input for MSMC2 was finalised using generate_multihetsep.py from msmc-tools, following the instructions in Schiffels and Wang (2020). After the generation of input data, two individuals from each population were randomly selected for analysis with MSMC2. Each individual was first analysed separately and then used in the cross-population analysis to infer the split time between the paired population. The results were visualised with R scripts (from msmc-tools) assuming a mutation rate of 1.42e-8 and a generation length of 2 years.

We used *moments* (Jouganous *et al.*, 2017) to infer the divergence time under two basic models: the isolation with migration model (IM) and strict isolation model (SI). We first used ANGSD v.0.921-3-g40ac3d6 (Korneliussen *et al.*, 2014) to estimate the folded site frequency spectrum (SFS) and then generated two-dimensional site frequency spectrums (2dSFS) for the selected population pairs. For each of the population pairs, the models were optimized in 10 independent runs, each consisting of 6 rounds of optimization and 30 replications for each optimization. We selected the best 10 replicates from each of the models according to the log-likelihood, and estimated the divergence times. The same mutation rate and generation length were used as in MSMC2 analyses. For details, see Table S2 and Script S1.

Simulated data from *msprime* (Kelleher *et al.*, 2016) was outputted in vcf format and used to replicate the analyses with *f4*-ratio test, MSMC2 and *moments*. For MSMC2 and *f4*-ratio tests, we followed the same pipeline as for empirical data. For *moments* analyses, we first use vcf2dadi.py script (available at https://github.com/CoBiG2/RAD_Tools/blob/master/vcf2dadi.py) to generate the 2dSFS for each of the population pair. The 2dSFSs were analysed with the same pipeline as the empirical data.

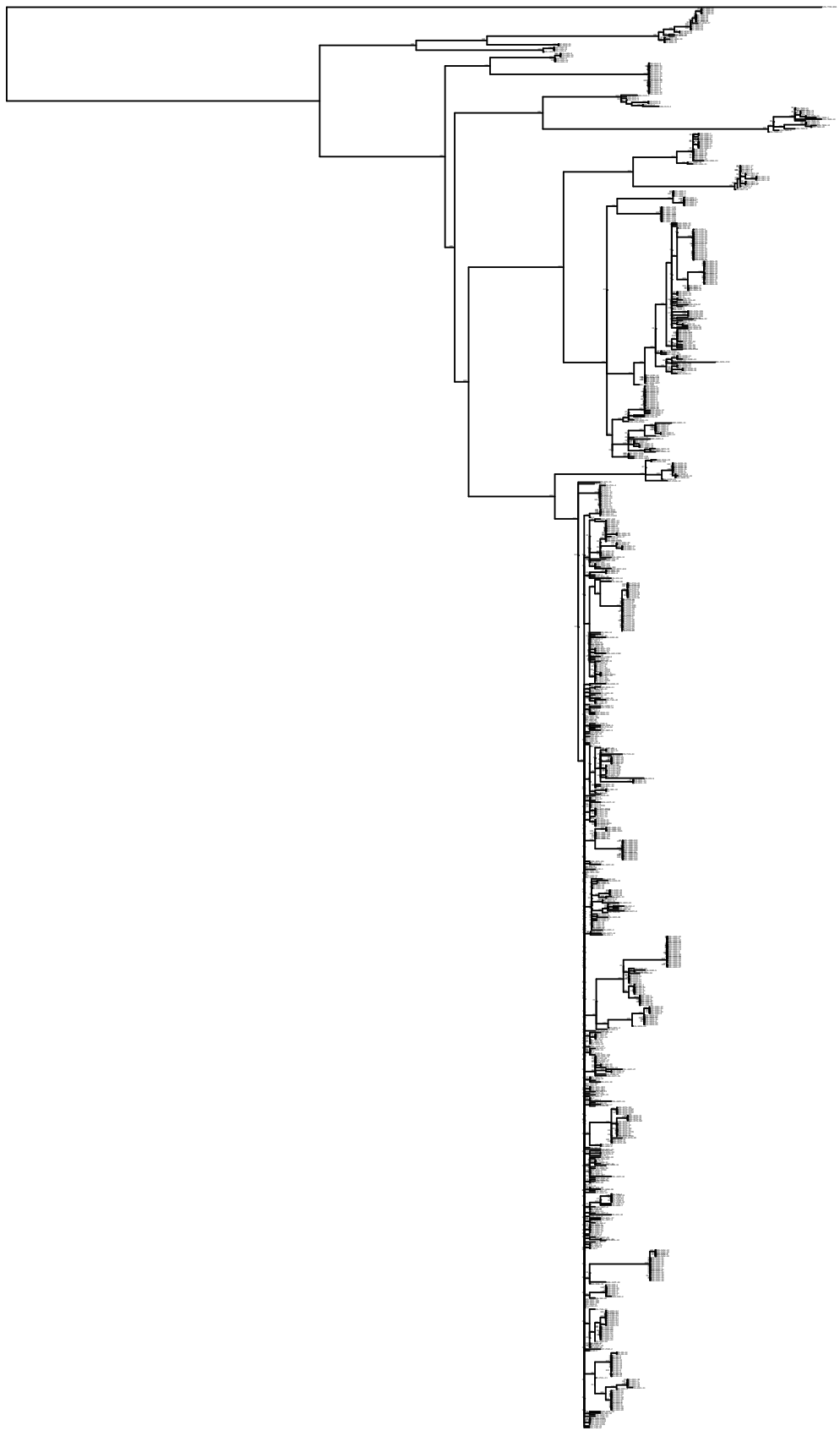

**FIG. S1.** The RAxML maximum likelihood tree for the mtDNA data with the taxon names and bootstrap support values.

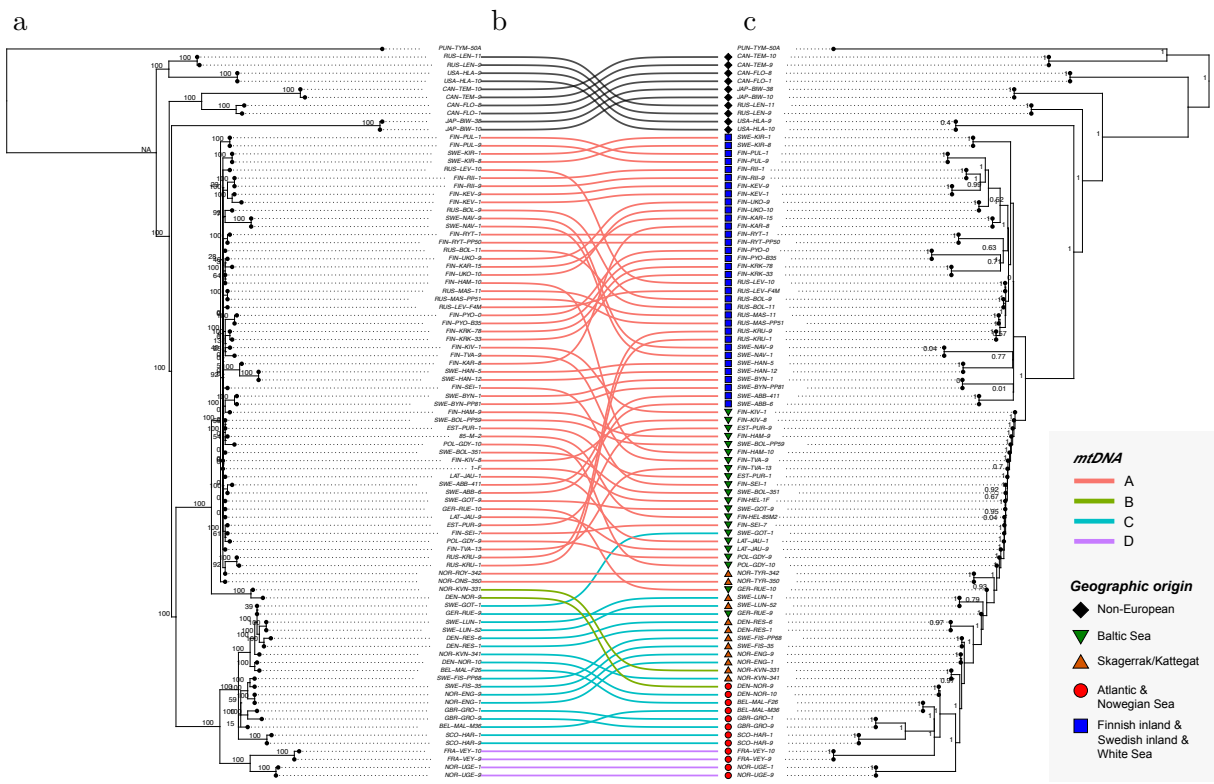


**FIG. S2.** Cytonuclear incongruence. a) The RAxML maximum likelihood tree for mtDNA data with taxon names and bootstrap support values. b) Lines, colored by the mitochondria type (cf. Fig. 1b), show incongruences between the two phylogenetic trees. c) The ASTRAL tree for autosomal DNA data with taxon names and local posterior probabilities. The colored shapes indicate the geographical origin of each sample and match those in Fig. 2b.

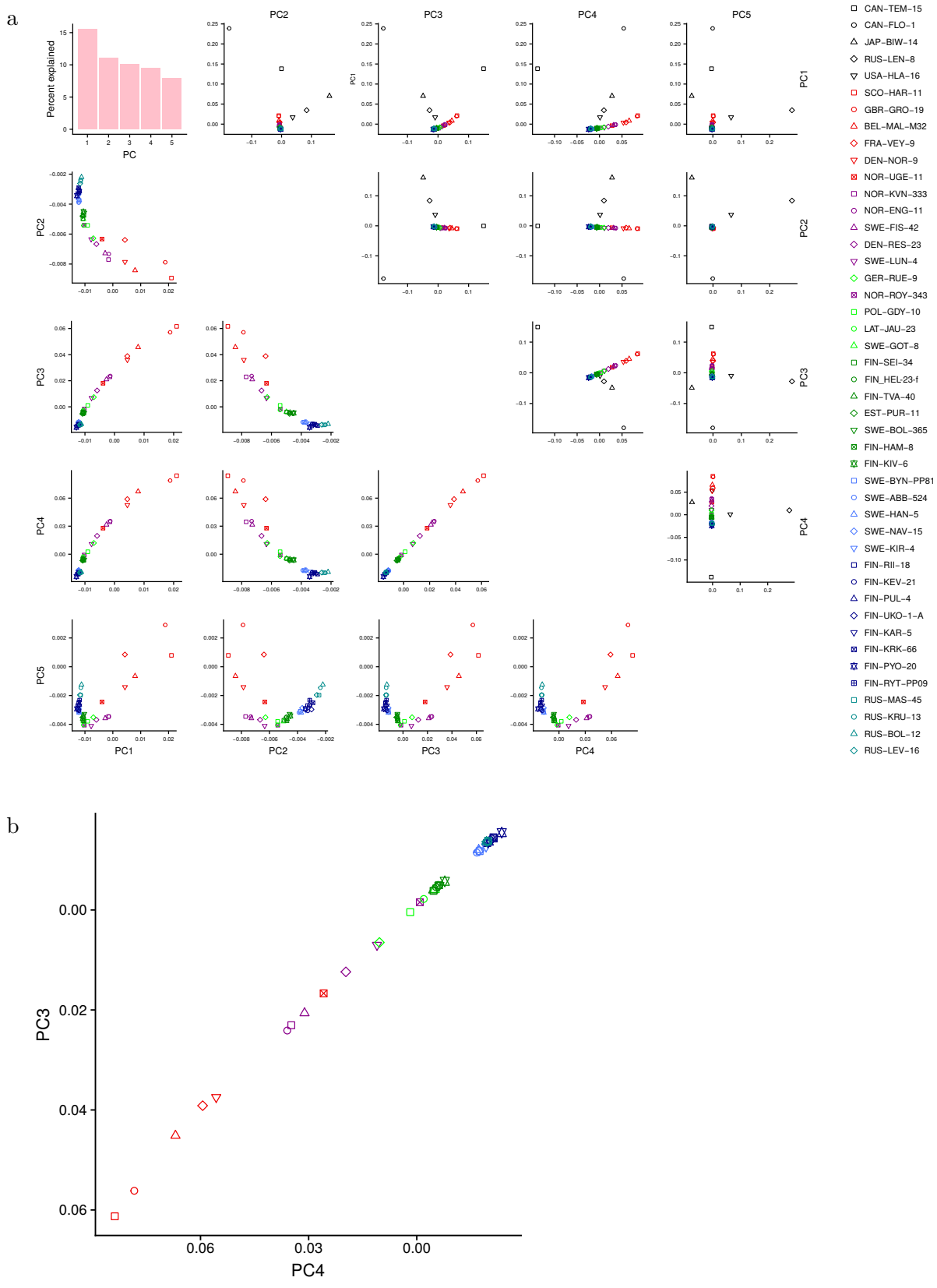





**FIG. S3.** Principal Component Analysis of the full nuclear variant data. a) The subplots show the first five PCA axes for the full data (top) and with the five non-European populations removed (bottom). The barplot shows the proportion of variation explained by each of the PCs. b) After removal of non-European populations, PC3 vs. PC4 shows a continuum from WL to EL with Kattegat/Skagerrak and Baltic populations in between. For clarity, one randomly selected individual from each population is shown.

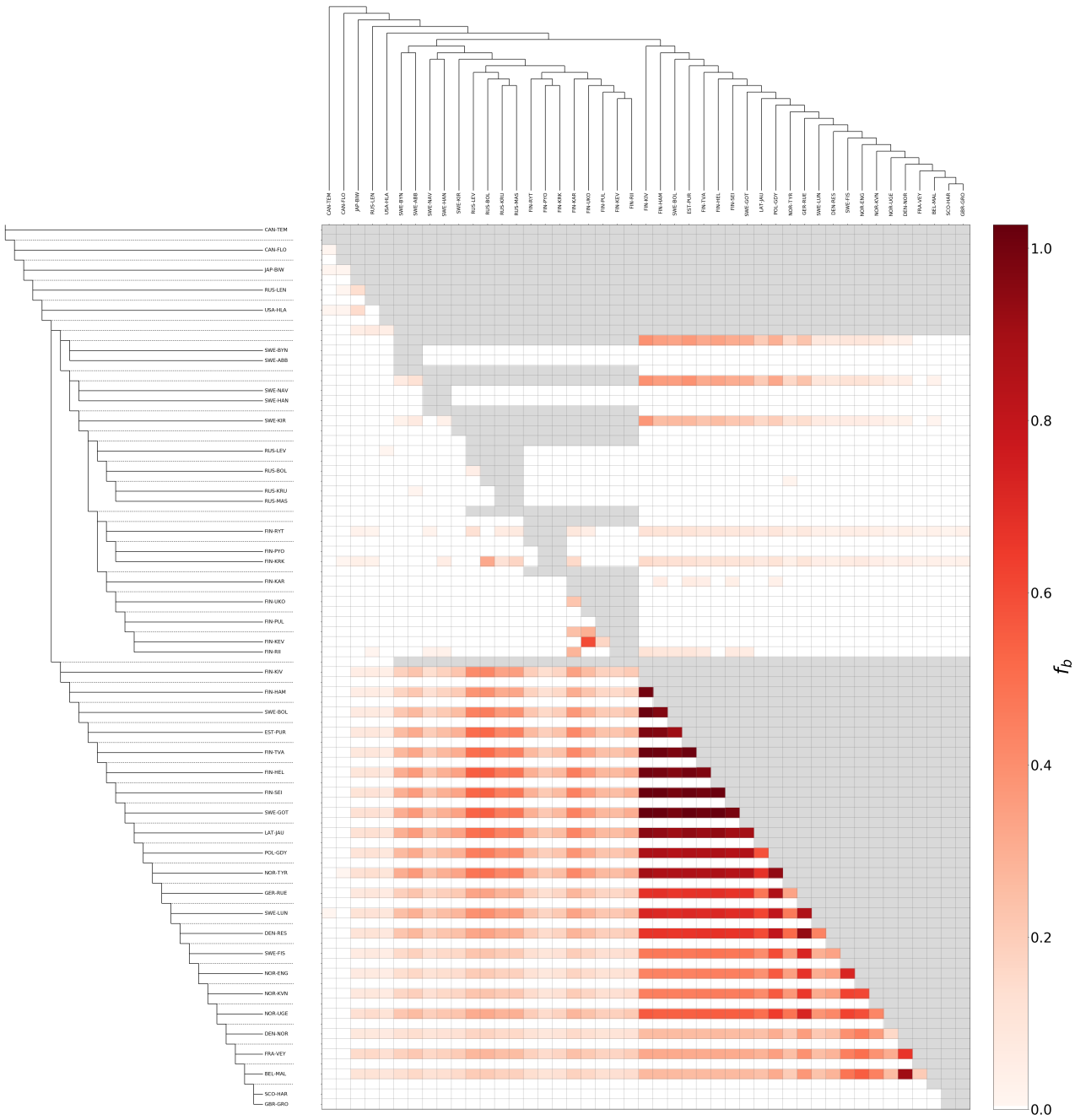


**FIG. S4.** The *f-branch* analysis identifies excess of shared derived alleles between populations (y axis) and the inferred species tree (x axis; ASTRAL tree with *Pungitius tymensis* as the outgroup). The population relationships are based on the frequency of "BBAA" patterns and the matrix values indicate excess of allele sharing between the branches on the y axis and the populations on the x axis, the higher fb values meaning greater allele sharing. All fb values are shown; values marked with gray cannot be estimated based on the given phylogeny.

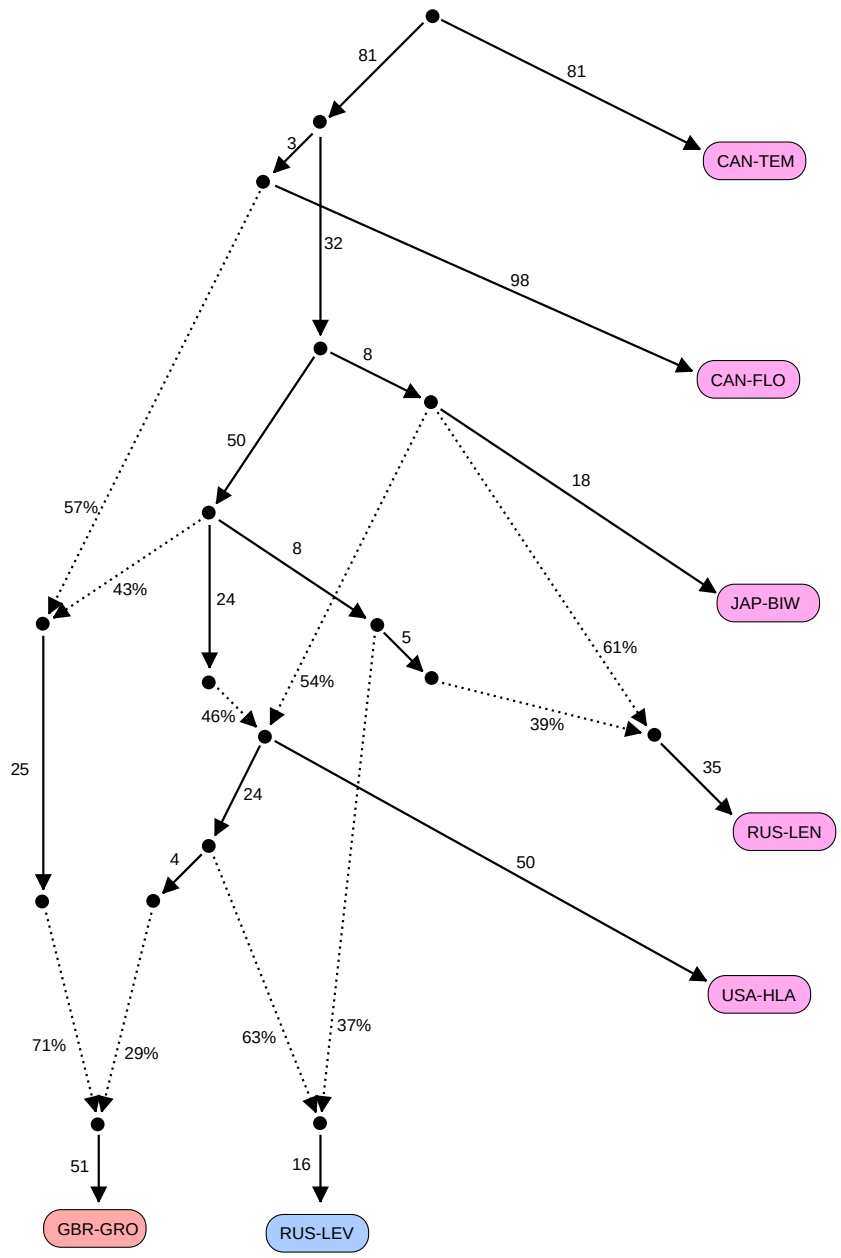


**FIG. S5.** History of North American and Asian populations and representative populations from Europe. The optimal qpGraph model (|Z|= 1.60) places the North American CAN-TEM and CAN-FLO and the Japanese JAP-BIW at the root. The freshwater populations from the River Lena (RUS-LEN) and Alaska (USA-HLA) derive, respectively, 61% and 54% of ancestry from an ancestor of the JAP-BIW, and 39% and 46% from populations representing the later waves of east-to-west migration. The EL population RUS-LEV derives its ancestry from the second and third waves of migration and shares one of these (the graph does not indicate which is earlier) with the WL population GBR-GRO. However, RUS-LEV (and the EL) lacks the old ancestry forming 40% (0.71 x 0.57 = 0.405) of the GBR-GRO genome.

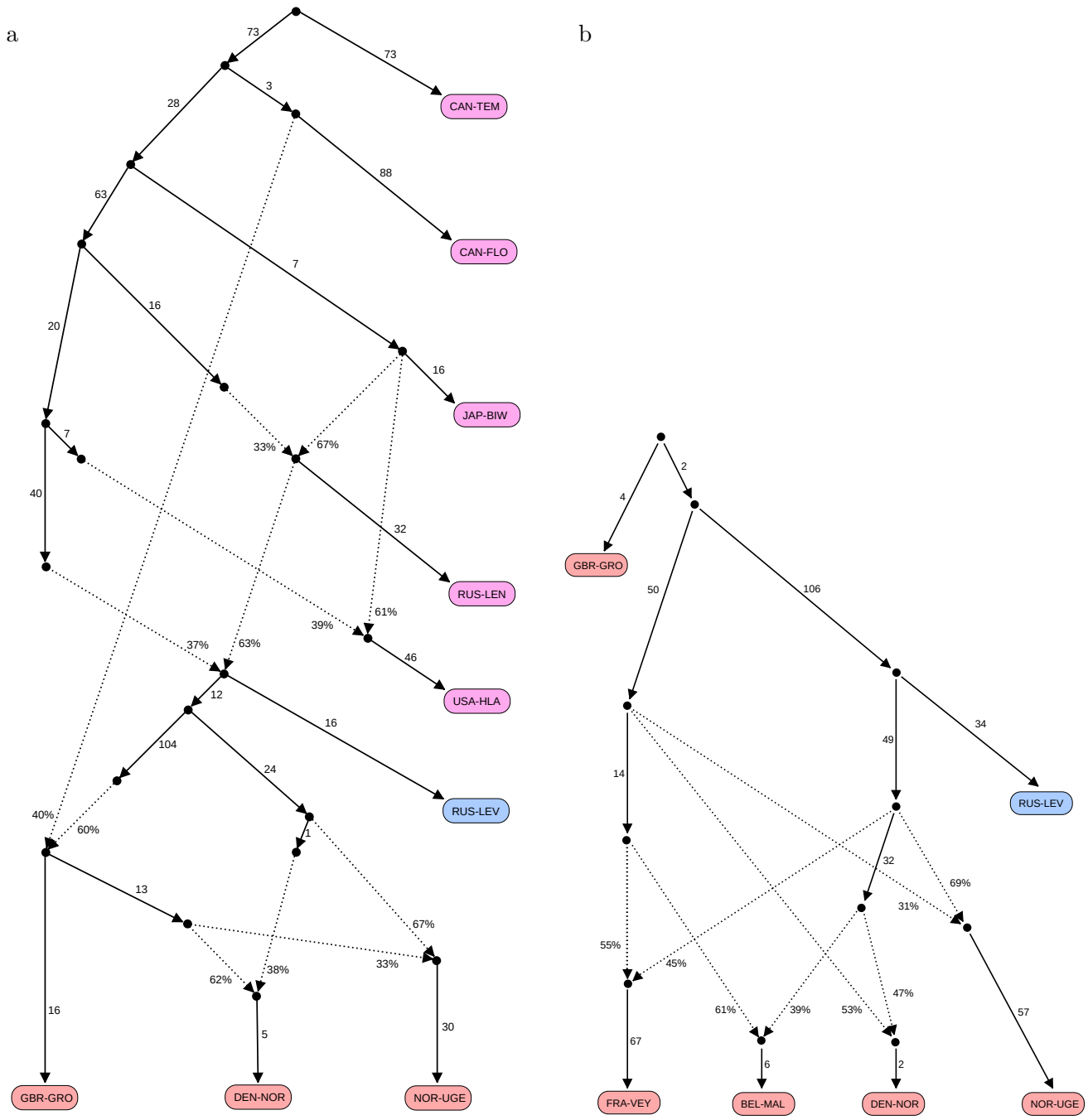


**FIG. S6.** The optimal qpGraph models for non-European populations (magenta) and representatives of WL and EL (red and blue, respectively). a) The full model ($|Z|= 2.60$) indicates a unique history for the isolated Norwegian population NOR-UGE. The Danish marine population DEN-NOR descends predominantly from the old WL (62%) but has a significant, more recent contribution from the east (38%). The proportions are almost identical to our basal model shown in Fig. 3. NOR-UGE descends predominantly from the east (67%) and has received a lower proportion (33%) from an old WL source shared with DEN-NOR. b) A submodel ($|Z|= 2.16$) for the European populations indicates a deep divergence of NOR-UGE and a longer shared history for the Belgian (BEL-MAL) and French (FRA-VEY) populations in comparison to DEN-NOR.

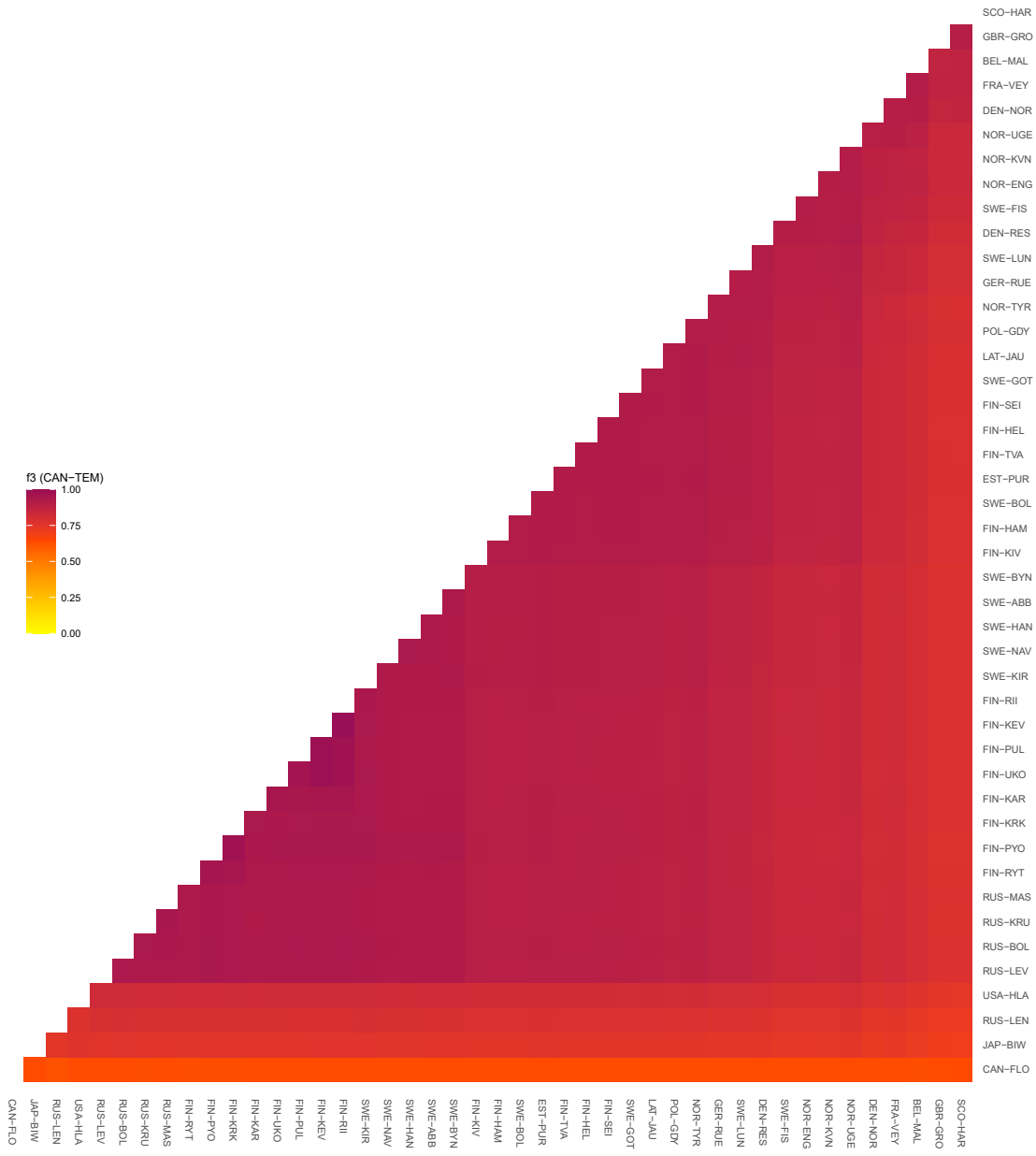


**FIG. S7.** The outgroup-*f3* analysis using CAN-TEM as the outgroup. The heatmap shows the statistics *f3*(X, Y; CAN-TEM) with darker values indicating a closer affinity between populations X and Y.

**FIG. S8.** Identification of optimal source pairs for the Baltic Sea populations using qpWave and qpAdm. a) Optimal source pairs for the Baltic Sea populations were identified using qpWave triplet analysis. RUS-LEV was fixed as representative for EL (Pop1). Each WL population (Pop2; y axis) was then tested as the "source" for each of the admixed populations (Pop3; x axis). The ANC5 set was used as the outgroup. The significance of prank1_tail values of each triplet is shown: triplets with insignificant results ($p>0.05$, red) mean that one of the populations in the triplet can be modeled as a mixture by the other two populations. b) Verification of triplets identified with qpWave using qpAdm. RUS-LEV was fixed as the EL source and the ANC5 set was used as the outgroup. Different WL populations were tested against the target populations. A significant result ($p<0.05$, shown in green) means that the source population pair (y axis) was rejected by the model and is not optimal for modelling the target population (x axis) as two-way admixture. The results indicate that DEN-NOR is the optimal source population for modeling the Baltic Sea populations. c) Admixture proportions with standard errors as estimated by qpAdm. RUS-LEV and DEN-NOR were used as the representatives of the EL and WL sources. Triangles and squares represent the ancestry from EL and WL, respectively. The blue and red color indicates if the model of two-way admixture for the given target population was rejected or not. The red dashed lines refers to admixture proportions of 0 and 1, and estimates beyond the range means the model being rejected. Note that the FRA-VEY population shows a high standard error: although the model was not rejected, it should be considered uncertain (Harney *et al.*, 2021).

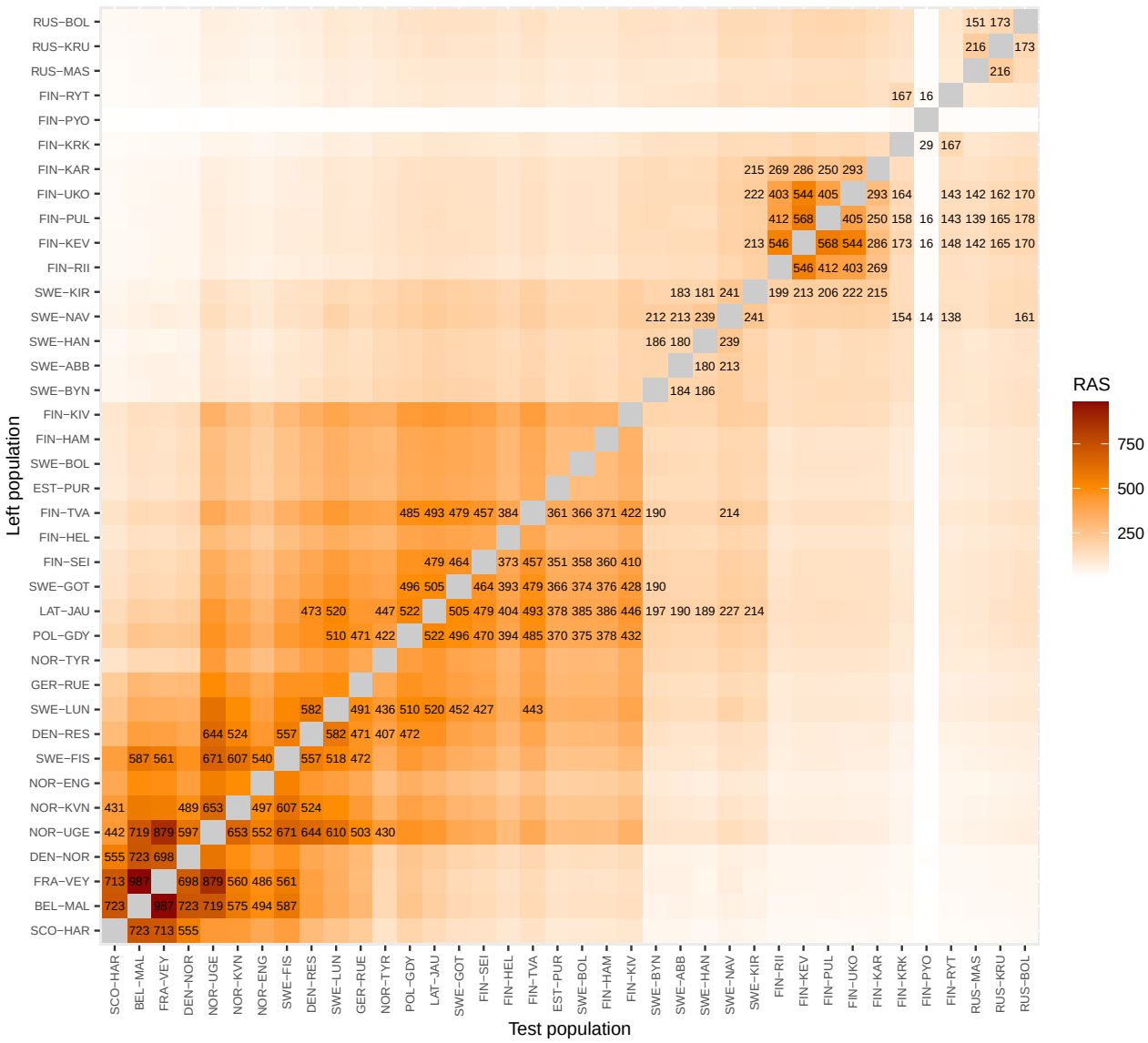


**FIG. S9.** Heatmap of rare alleles sharing statistics (RASS) for each test population. Rare alleles with allele count between 2-10 in the reference set (ANC5, GBR-GRO and RUS-LEV) were included in the statistics. Alleles were polarised using *Pungitius tymensis*. The five highest RASS values for each test population are shown.

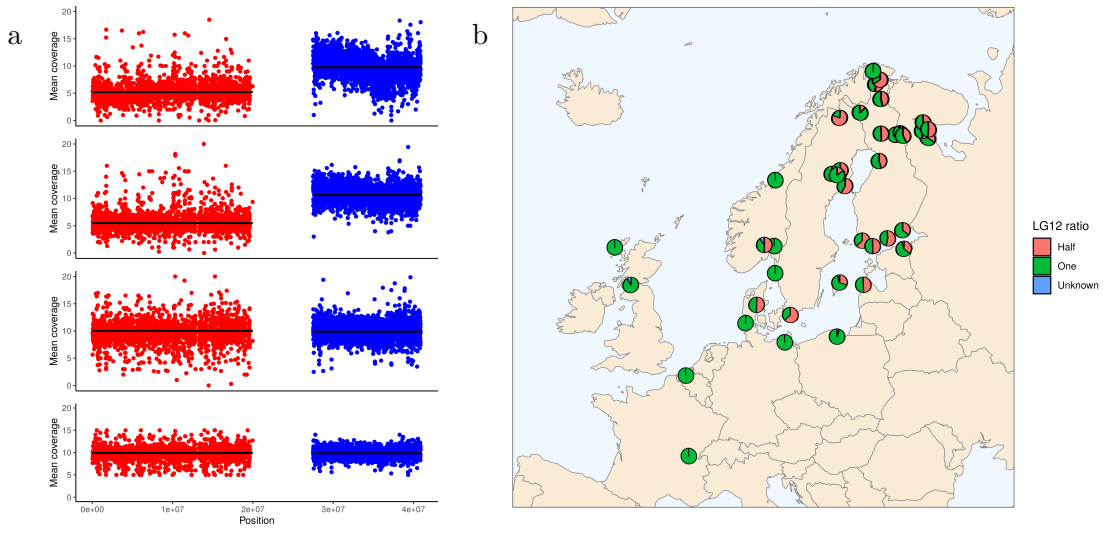


**FIG. S10.** a) The mean sequencing coverage across the sex-chromosome (red) and pseudo-autosomal (blue) parts of LG12 in two FIN-PYO males (top) and females (bottom) using 5 kbp windows. The black lines show the average for each region. b) The proportion of samples showing the ratio of half or one for the sequencing coverage across the sex-chromosome and pseudo-autosomal regions. The ratio of half indicates EL male (i.e., EL Y-chromosome) while the ratio of one can either indicate EL female (X-chromosome) or a different sex chromosome system.

**Table S1.** Sample information, mtDNA type (A-D) and lineage assignment.

| This study (WGS) | | | | | | | | | | | Teacher et al. 2011 (mtDNA) | | Guo et al. 2019 (RAD-seq) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | mtDNA type | | | | | | | | | |
| Label | Lat | Long | Habitat | Region | N | A | B | C | D | Assignment | Label | Assignment | Label | Assignment | Species |
| BEL-MAL | 51.17 | 3.47 | River | Central Belgium | 20 | 0 | 0 | 20 | 0 | WL | BE-LEU | WE | PP-BE-MAL | WE | *P.pungitius* |
| DEN-NOR | 54.98 | 8.66 | Marine | North Sea | 25 | 0 | 11 | 14 | 0 | WL | | | | | *P.pungitius* |
| DEN-RES | 56.18 | 9.63 | Pond | Central Denmark | 20 | 0 | 0 | 20 | 0 | WL | DK-GUD | WE | | | *P.pungitius* |
| EST-PUR | 59.42 | 26.98 | River | Coastal Baltic Sea | 19 | 18 | 1 | 0 | 0 | EL | EE-PUR | EE | | | *P.pungitius* |
| FIN-HAM | 60.56 | 27.19 | Marine | Baltic Sea | 20 | 20 | 0 | 0 | 0 | EL | | | | | *P.pungitius* |
| FIN-HEL | 60.2 | 25.18 | Marine | Baltic Sea | 22 | 22 | 0 | 0 | 0 | EL | FI-HEL | EE | PP-FI-HEL | EE | *P.pungitius* |
| FIN-KAR | 66.66 | 26.44 | Pond | Northern Finland | 20 | 20 | 0 | 0 | 0 | EL | FI-KAR | EE | | | *P.pungitius* |
| FIN-KEV | 69.75 | 27.02 | River | Northern Finland | 20 | 20 | 0 | 0 | 0 | EL | FI-KEV | EE | | | *P.pungitius* |
| FIN-KIV | 65 | 25.47 | Marine | Baltic Sea | 19 | 19 | 0 | 0 | 0 | EL | FI-KIV | EE | | | *P.pungitius* |
| FIN-KRK | 66.44 | 29.14 | Pond | Northern Finland | 20 | 20 | 0 | 0 | 0 | EL | FI-KRK | EE | | | *P.pungitius* |
| FIN-PUL | 69.97 | 27.97 | Lake | Northern Finland | 17 | 17 | 0 | 0 | 0 | EL | FI-PUL | EE | | | *P.pungitius* |
| FIN-PYO | 66.26 | 29.43 | Pond | Northern Finland | 31 | 31 | 0 | 0 | 0 | EL | FI-PYO | EE | | | *P.pungitius* |
| FIN-RII | 68.11 | 23.57 | Lake | Northern Finland | 23 | 23 | 0 | 0 | 0 | EL | FI-RIK | EE | | | *P.pungitius* |
| FIN-RYT | 66.38 | 29.32 | Pond | Northern Finland | 21 | 21 | 0 | 0 | 0 | EL | FI-RYT | EE | PP-FI-RYT | EE | *P.pungitius* |
| FIN-SEI | 60.23 | 21.95 | Marine | Baltic Sea | 24 | 24 | 0 | 0 | 0 | EL | | | | | *P.pungitius* |
| FIN-TVA | 59.83 | 23.2 | Marine | Baltic Sea | 22 | 22 | 0 | 0 | 0 | EL | | | | | *P.pungitius* |
| FIN-UKO | 68.78 | 27.44 | Lake | Northern Finland | 22 | 22 | 0 | 0 | 0 | EL | | | | | *P.pungitius* |
| FRA-VEY | 46.23 | 5.12 | River | Central France | 16 | 0 | 0 | 0 | 16 | EL | | | | | *P.pungitius* |
| GBR-GRO | 57.62 | -7.51 | River | British Isles | 19 | 0 | 0 | 19 | 0 | WL | | | | | *P.pungitius* |
| GER-RUE | 54.01 | 13 | Marine | Baltic Sea | 27 | 16 | 1 | 10 | 0 | WL+EL | GE-RUG | WE+EE | | | *P.pungitius* |
| LAT-JAU | 57.51 | 21.69 | River | Coastal Baltic Sea | 21 | 19 | 0 | 2 | 0 | EL | | | | | *P.pungitius* |
| NOR-ENG | 59.9 | 10.53 | Lake | Skagerrak | 23 | 0 | 0 | 23 | 0 | WL | NO-ENG | WE | | | *P.pungitius* |
| NOR-KVN | 70.58 | 26.99 | Lake | Barents Sea coast | 11 | 0 | 1 | 10 | 0 | WL | | | | | *P.pungitius* |
| NOR-TYR | 59.91 | 10.3 | Lake | Skagerrak | 8 | 8 | 0 | 0 | 0 | EL | | | | | *P.pungitius* |
| NOR-UGE | 63.96 | 10.43 | Lake | Coastal Norwegian Sea | 20 | 0 | 0 | 0 | 20 | WL | NO-UGE | WE | PP-NO-UGE | WE | *P.pungitius* |
| POL-GDY | 54.4 | 18.53 | Marine | Baltic Sea | 20 | 20 | 0 | 0 | 0 | WL+EL | PL-GDY | EE | | | *P.pungitius* |
| RUS-BOL | 66.3 | 33.4 | Pond | White Sea | 20 | 20 | 0 | 0 | 0 | EL | RU-BOL | EE | PS-RU-BOL | EE | *P.pungitius* |
| RUS-KRU | 66.3 | 33.4 | Pond | White Sea | 20 | 20 | 0 | 0 | 0 | EL | RU-KRU | EE | | | *P.pungitius* |
| RUS-LEV | 66.3 | 33.4 | Marine | White Sea | 30 | 30 | 0 | 0 | 0 | EL | RU-LEV | EE | PP-RU-LEV | EE | *P.pungitius* |
| RUS-MAS | 66.3 | 33.4 | Pond | White Sea | 21 | 21 | 0 | 0 | 0 | EL | RU-MAS | EE | | | *P.pungitius* |
| SCO-HAR | 55.75 | -4.42 | Lake | British Isles | 10 | 0 | 0 | 10 | 0 | WL | UK-HAR | WE | PP-UK-HAR | WE | *P.pungitius* |
| SWE-ABB | 64.48 | 19.44 | Pond | Swedish Inland | 21 | 21 | 0 | 0 | 0 | EL | SE-ABB | EE | | | *P.pungitius* |
| SWE-BOL | 63.66 | 20.21 | Marine | Baltic Sea | 21 | 21 | 0 | 0 | 0 | EL | SE-BOL | EE | | | *P.pungitius* |
| SWE-BYN | 64.46 | 19.44 | Pond | Swedish Inland | 23 | 23 | 0 | 0 | 0 | EL | SE-BYN | EE | PP-SE-BYN | EE | *P.pungitius* |
| SWE-FIS | 58.23 | 11.4 | Marine | Kattegat | 20 | 1 | 0 | 19 | 0 | WL+EL | SE-FIS | WE+EE | PP-SE-FIS | WE | *P.pungitius* |
| SWE-GOT | 57.73 | 18.95 | Marine | Baltic Sea | 19 | 18 | 0 | 1 | 0 | WL+EL | | | | | *P.pungitius* |
| SWE-HAN | 64.56 | 19.17 | Pond | Swedish Inland | 20 | 20 | 0 | 0 | 0 | EL | SE-HAN | EE | | | *P.pungitius* |
| SWE-KIR | 67.9 | 20.09 | Pond | Swedish Inland | 15 | 15 | 0 | 0 | 0 | EL | | | | | *P.pungitius* |
| SWE-LUN | 55.72 | 13.43 | Pond | Baltic Sea (isolated) | 21 | 0 | 0 | 21 | 0 | WL+EL | SE-LUN | WE | | | *P.pungitius* |
| SWE-NAV | 64.56 | 19.2 | Pond | Swedish Inland | 20 | 20 | 0 | 0 | 0 | EL | SE-NAV | EE | | | *P.pungitius* |
| CAN-FLO | 54.23 | -111.63 | Lake | Canadian Inland | 8 | | | | | ANC | | | PP-CA-FLO | NA | *P.pungitius* |
| CAN-TEM | 47.71 | -68.91 | Lake | Canadian Inland | 16 | | | | | ANC | CA-TAM | CA(NA) | PP-CA-TEM | NA | *P.pungitius* |
| JAP-BIW | 43.08 | 145.11 | Marine | Japan Sea | 24 | | | | | ANC | | | PP-JP-BIW(BW) | FE | *P.pungitius* |
| RUS-LEN | 72.98 | 126.96 | River | Fareast | 10 | | | | | ANC | | | PP-RU-LEN | FE | *P.pungitius* |
| USA-HLA | 61.59 | -149.76 | Lake | Alaskan | 20 | | | | | ANC | | | | | *P.pungitius* |
| PUN-TYM | 43.82 | 145.08 | River | Japan inland | 1 | | | | | Outgroup | | | PT-JP-MOT | Outgroup | *P.tymensis* |

**Table S2.** Data filtering for different analyses.

| Name | Filtering criteria | SNPs | Analyses used |
|---|---|---|---|
| SNP Set 0 | Of original dataset, keep autosomal biallelic SNPs in non-repetitive regions; remove interspecific variants | | *NA* |
| SNP Set 1 | From SNP Set 0, keep only binary SNPs with quality score ≥30, mean coverage ≥8x, GQ ≥ 20 | 152,175,77 | *NA* |
| SNP Set 2 | From SNP Set 1, filter in vcftools with –minGQ 20 –minQ 30 –min-meanDP 8 –max-meanDP 25 –max-missing 0.9, and then LD pruning with plink –indep-pairwise 50 10 0.1 | | PCA, ADMIXTURE |
| SNP Set 3 | From SNP Set 1, select two random samples from each populations allowing maximum 50% missing data and keep only variable sites with bcftools view (–min-ac=1) | | RAxML and ASTRAL |
| SNP Set 4 | From SNP Set 1, keep only variable sites with maximum mean DP of 25 and no more than 25% missing data; distance between SNPs required to be at least 300bp | 506,171 | ADMIXTOOLS (D statistic, outgroup-*f3*, *f4*-ratio, qpWave, qpAdm and qpGraph test) and RASS |
| SNP Set 5 | From SNP Set 1, keep only variable sites with maximum mean DP of 25 and no more than 10% missing data | | *f-branch* |
| | For details, see https://github.com/XueyunF/nsp_phylogeo/blob/main/Divergence_time/MSMC2_pipeline.md | | MSMC2 |
| | For details, see https://github.com/XueyunF/nsp_phylogeo/blob/main/Divergence_time/Moments_pipeline.md | | *moments* |
| SNP Set 6 | Subsets of the EL sex chromosome (LG12). For details, see https://github.com/XueyunF/nsp_phylogeo/blob/main/Sex_chromosome/README.md | | EL sex determination |
| SNP Set 7 | Complete mitochondrial genome. For details, see https://github.com/XueyunF/nsp_phylogeo/tree/main/Structure_and_Phylogeny | | RAxML |

**Table S3.** Detailed results of qpWave analysis: WL, EL and Baltic Sea populations.

| | Left populations | Right populations (outgroups) | p_rnk 0_tail | p_rnk 1_tail | p_rnk 2_tail | p_rnk 3_tail | p_rnk 4_tail | p_rnk 5_tail | p_rnk 6_tail | p_rnk 0_tldf | p_rnk 1_tldf | p_rnk 2_tldf | p_rnk 3_tldf | p_rnk 4_tldf | p_rnk 5_tldf | p_rnk 6_tldf |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Test WL | GBR-GRO, SCO-HAR, DEN-NOR, FRA-VEY, NOR-UGE, BEL-MAL, NOR-KVN, SWE-FIS, NOR-ENG | CAN-TEM, CAN-FLO, JAP-BIW, RUS-LEN, USA-HLA, RUS-LEV | 0 | 2.86E-60 | 0.209 | 0.681 | 0.89 | 1 | NA | 1 | 0 | 1.93E-67 | 0.058 | 0.387 | 0.89 | NA |
| Test EL | RUS-LEV, RUS-KRU, RUS-MAS, RUS-BOL, FIN-KRK, FIN-PYO, FIN-RYT, FIN-UKO, FIN-RII, FIN-PUL, FIN-KEV, FIN-KAR, SWE-ABB, SWE-HAN, SWE-BYN, SWE-NAV, SWE-KIR | CAN-TEM, CAN-FLO, JAP-BIW, RUS-LEN, USA-HLA, GBR-GRO | 7.98E-94 | 3.33E-13 | 0.003 | 0.829 | 0.998 | 1 | NA | 1 | 5.3E-92 | 4.05E-14 | 1.11E-05 | 0.282 | 0.998 | NA |
| Test BS | FIN-KIV, FIN-HAM, SWE-BOL, EST-PUR, FIN-TVA, FIN-HEL, FIN-SEI, SWE-GOT, LAT-JAU, POL-GDY, GER-RUE, SWE-LUN | CAN-TEM, CAN-FLO, JAP-BIW, RUS-LEN, USA-HLA, GBR-GRO, RUS-LEV | 0 | 2.56E-66 | 0.071 | 0.891 | 0.966 | 0.835 | 1 | 1 | 0 | 1.93E-77 | 0.000902 | 0.45 | 0.918 | 0.835 |
| Test BS* | FIN-KIV, FIN-HAM, SWE-BOL, EST-PUR, FIN-TVA, FIN-HEL, FIN-SEI | CAN-TEM, CAN-FLO, JAP-BIW, RUS-LEN, USA-HLA, GBR-GRO, RUS-LEV | 4.4E-162 | 0.144 | 0.891 | 0.878 | 0.993 | 0.791 | 1 | 1 | 2.4E-175 | 0.00625 | 0.655 | 0.518 | 0.981 | 0.791 |

*Excluding southern Baltic Sea populations and LAT-JAU

**Table S4.** Detailed results of qpWave analysis: WL and EL populations.

| | Left populations | | Results | | | | | | Right populations (outgroups) |
|---|---|---|---|---|---|---|---|---|---|
| | FIX | TEST | p_rank0 _tail | p_rank1 _tail | p_rank2 _tail | p_rank0 _taildiff | p_rank1 _taildiff | p_rank2 _taildiff | |
| Within WL | GBR-GRO, SCO-HAR | DEN-NOR | <0.001 | 0.0239 | 1 | 1 | <0.001 | 0.0239 | ANC5, RUS-LEV |
| | | BEL-MAL | <0.001 | 0.07 | 1 | 1 | <0.001 | 0.07 | |
| | | FRA-VEY | <0.001 | 0.0885 | 1 | 1 | <0.001 | 0.0885 | |
| | | NOR-UGE | 0 | 0.0222 | 1 | 1 | 0 | 0.0222 | |
| | | NOR-KVN | <0.001 | 0.001 | 1 | 1 | <0.001 | 0.001 | |
| | | SWE-FIS | 0 | 0.004 | 1 | 1 | 0 | 0.00398 | |
| | | NOR-ENG | 0 | 0.009 | 1 | 1 | 0 | 0.009 | |
| | | DEN-RES | 0 | 0.004 | 1 | 1 | 0 | 0.004 | |
| Within EL | RUS-LEV, RUS-BOL | RUS-KRU | 0.151 | 0.152 | 1 | 1 | 0.254 | 0.152 | ANC5, GBR-GRO |
| | | RUS-MAS | 0.595 | 0.628 | 1 | 1 | 0.451 | 0.628 | |
| | | FIN-RYT | 0.891 | 0.984 | 1 | 1 | 0.594 | 0.984 | |
| | | FIN-PYO | 0.802 | 0.867 | 1 | 1 | 0.559 | 0.867 | |
| | | FIN-KRK | 0.507 | 0.356 | 1 | 1 | 0.56 | 0.356 | |
| | | FIN-UKO | 0.182 | 0.411 | 1 | 1 | 0.132 | 0.411 | |
| | | FIN-PUL | 0.07 | 0.413 | 1 | 1 | 0.0392 | 0.413 | |
| | | FIN-KEV | 0.0249 | 0.418 | 1 | 1 | 0.0109 | 0.418 | |
| | | FIN-KAR | <0.001 | 0.407 | 1 | 1 | <0.001 | 0.407 | |
| | | SWE-ABB | <0.001 | 0.442 | 1 | 1 | <0.001 | 0.442 | |
| | | SWE-HAN | <0.001 | 0.418 | 1 | 1 | <0.001 | 0.418 | |
| | | SWE-BYN | <0.001 | 0.392 | 1 | 1 | <0.001 | 0.392 | |
| | | SWE-NAV | <0.001 | 0.441 | 1 | 1 | <0.001 | 0.441 | |
| | | SWE-KIR | 0.02 | 0.373 | 1 | 1 | 0.01 | 0.373 | |

**Table S5.** Detailed results of qpWave analysis: WL populations.

| Left populations | | Results | | | | | | | | | | Right populations (outgroups) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| FIX | TEST | p_rank0 _tail | p_rank1 _tail | p_rank2 _tail | p_rank3 _tail | p_rank4 _tail | p_rank0 _taildiff | p_rank1 _taildiff | p_rank2 _taildiff | p_rank3 _taildiff | p_rank4 _taildiff | |
| GBR-GRO, SCO-HAR, BEL-MAL, FRA-VEY | - | <0.001 | 0.228 | 0.713 | 1 | | 1 | <0.001 | 0.102 | 0.713 | | ANC5, RUS-LEV |
| | DEN-NOR | <0.001 | <0.001 | 0.948 | 0.9 | 1 | 1 | <0.001 | <0.001 | 0.832 | 0.908 | |
| | NOR-UGE | 0 | <0.001 | 0.82 | 0.785 | 1 | 1 | 0 | <0.001 | 0.658 | 0.785 | |
| | NOR-KVN | 0 | <0.001 | 0.843 | 0.965 | 1 | 1 | 0 | <0.001 | 0.618 | 0.965 | |
| | SWE-FIS | 0 | <0.001 | 0.877 | 0.891 | 1 | 1 | 0 | <0.001 | 0.7 | 0.891 | |
| | NOR-ENG | 0 | <0.001 | 0.846 | 0.969 | 1 | 1 | 0 | <0.001 | 0.621 | 0.969 | |
| | DEN-RES | 0 | <0.001 | 0.844 | 0.892 | 1 | 1 | 0 | <0.001 | 0.648 | 0.892 | |
| GBR-GRO, SCO-HAR, BEL-MAL, FRA-VEY | - | <0.001 | 0.744 | 0.973 | 1 | NA | 1 | <0.001 | 0.486 | 0.973 | NA | ANC5 |
| | DEN-NOR | <0.001 | 0.006 | 0.972 | 0.699 | 1 | 1 | <0.001 | <0.001 | 0.948 | 0.699 | |
| | NOR-UGE | <0.001 | 0.0526 | 0.801 | 0.811 | 1 | 1 | <0.001 | 0.01 | 0.663 | 0.811 | |
| | NOR-KVN | <0.001 | 0.0263 | 0.941 | 0.898 | 1 | 1 | <0.001 | 0.00284 | 0.857 | 0.898 | |
| | SWE-FIS | <0.001 | <0.001 | 0.976 | 0.685 | 1 | 1 | <0.001 | <0.001 | 0.958 | 0.685 | |
| | NOR-ENG | <0.001 | 0.007 | 0.945 | 0.857 | 1 | 1 | <0.001 | <0.001 | 0.869 | 0.857 | |
| | DEN-RES | <0.001 | <0.001 | 0.839 | 0.862 | 1 | 1 | <0.001 | <0.001 | 0.706 | 0.862 | |

**Table S6.** Two population symmetrical test with qpWave, the outgroups was set as ANC5.

Table S6 is available at https://github.com/XueyunF/nsp_phylogeo/blob/main/Tables/Two_pop_symmetrical_qpwave.txt

**Table S7.** Detailed results qpWave testing of triples for the Baltic sea populations.

Table S7 is available at https://github.com/XueyunF/nsp_phylogeo/blob/main/Tables/Baltic_triplets_qpWave.txt

**Table S8.** Detailed results of *f4*-ratio test with the EL as source.

Table S8 is available at https://github.com/XueyunF/nsp_phylogeo/blob/main/Tables/f4_ratio_sourceEL.txt

**Table S9.** Detailed results of *f4*-ratio test with the WL as source.

Table S9 is available at https://github.com/XueyunF/nsp_phylogeo/blob/main/Tables/f4_ratio_sourceWL.txt

**Table S10.** Detailed results of D-statistics.
Table S10 is available at https://github.com/XueyunF/nsp_phylogeo/blob/main/Tables/Dstat_result_full_table.txt.

**Script S1.** Python-code for demographic analyses with *moments*

The code is available at https://github.com/XueyunF/nsp_phylogeo/tree/main/Divergence_time.

**Script S2.** Python-code for data simulation with *msprime*

The code is available at https://github.com/XueyunF/nsp_phylogeo/tree/main/Simulations.