PLOS GENETICS

RESEARCH ARTICLE

The genomic landscape at a late stage of stickleback speciation: High genomic divergence interspersed by small localized regions of introgression

Mark Ravinet[1,2], Kohta Yoshida[1,3], Shuji Shigenobu[4], Atsushi Toyoda[5], Asao Fujiyama[5], Jun Kitano[1]*

1 Division of Ecological Genetics, Department of Population Genetics, National Institute of Genetics, Mishima, Shizuoka, Japan, **2** Centre for Ecological and Evolutionary Synthesis, University of Oslo, Oslo, Norway, **3** Integrative Evolutionary Biology, Max Planck Institute for Developmental Biology, Tübingen, Germany, **4** Functional Genomics Facility, National Institute for Basic Biology, Okazaki, Aichi, Japan, **5** Comparative Genomics Laboratory, National Institute of Genetics, Mishima, Shizuoka, Japan

* jkitano@nig.ac.jp

Citation: Ravinet M, Yoshida K, Shigenobu S, Toyoda A, Fujiyama A, Kitano J (2018) The genomic landscape at a late stage of stickleback speciation: High genomic divergence interspersed by small localized regions of introgression. PLoS Genet 14(5): e1007358. https://doi.org/10.1371/journal.pgen.1007358

Editor: Bret A. Payseur, University of Wisconsin–Madison, UNITED STATES

Received: July 27, 2017

Accepted: April 11, 2018

Published: May 23, 2018

Data Availability Statement: All sequence data are deposited in DDBJ (DRA005130 and DRA006702). Custom R scripts are available from Dryad (doi:10.5061/dryad.104g3d0)

Funding: MR and KY were supported by a Standard Postdoctoral Research Fellowship for Research in Japan and a Postdoctoral Research Fellowship for Young Scientists PD from JSPS, respectively. This research was supported by NIG collaborative grant –I (98I2017) to MR, Grant-in-

Abstract

Speciation is a continuous process and analysis of species pairs at different stages of divergence provides insight into how it unfolds. Previous genomic studies on young species pairs have revealed peaks of divergence and heterogeneous genomic differentiation. Yet less known is how localised peaks of differentiation progress to genome-wide divergence during the later stages of speciation in the presence of persistent gene flow. Spanning the speciation continuum, stickleback species pairs are ideal for investigating how genomic divergence builds up during speciation. However, attention has largely focused on young postglacial species pairs, with little knowledge of the genomic signatures of divergence and introgression in older stickleback systems. The Japanese stickleback species pair, composed of the Pacific Ocean three-spined stickleback (*Gasterosteus aculeatus*) and the Japan Sea stickleback (*G. nipponicus*), which co-occur in the Japanese islands, is at a late stage of speciation. Divergence likely started well before the end of the last glacial period and crosses between Japan Sea females and Pacific Ocean males result in hybrid male sterility. Here we use coalescent analyses and Approximate Bayesian Computation to show that the two species split approximately 0.68–1 million years ago but that they have continued to exchange genes at a low rate throughout divergence. Population genomic data revealed that, despite gene flow, a high level of genomic differentiation is maintained across the majority of the genome. However, we identified multiple, small regions of introgression, occurring mainly in areas of low recombination rate. Our results demonstrate that a high level of genome-wide divergence can establish in the face of persistent introgression and that gene flow can be localized to small genomic regions at the later stages of speciation with gene flow.

Aid for Scientific Research on Innovative Areas "Gene Correlative System" (23113007 and 23113001) to JK, "Genome Science" (221S0002) to AT and AF, and JSPS KAKENHI (15H02418) to JK, NIBB Collaborative Research Program (10-337, 11-311) to JK. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Author summary

When species evolve, reproductive isolation leads to a build-up of differentiation in the genome where genes involved in the process occur. Spanning the speciation continuum, stickleback species pairs are ideal for investigating how genomic divergence accumulates during speciation. However, much of our understanding of stickleback speciation comes from early stage divergence, with relatively few examples from more divergent species pairs that still exchange genes. To address this, we focused on Pacific Ocean and Japan Sea sticklebacks, which co-occur in the Japanese islands. We established that they are the oldest and most divergent known stickleback species pair, that they evolved in the face of gene flow and that this gene flow is still on going. We found introgression is confined to small, localised genomic regions where recombination rate is high. Our results show high divergence can be maintained between species, despite extensive gene flow.

Introduction

Speciation is a continuous process through which reproductive isolation is established [1–3]. According to the genic view of speciation [4], when populations are in contact, gene flow is initially restricted at barrier loci (i.e. loci underlying reproductive isolation), leading to the emergence of peaks of genetic differentiation surrounding such barriers; i.e. heterogeneous genomic differentiation [5,6]. As speciation progresses, this localised build-up of reproductive isolation spreads to nearby regions due to linkage disequilibrium [4,5,7]. Once a critical amount of differentiation at multiple barrier loci has accumulated, reduction of the genome-wide effective migration rate will eventually lead to divergence across the entire genome [5,7]. This final step of genome-wide congealing may be a rapid and non-linear phase transition under certain conditions, such as when isolating barriers have a polygenic basis or a few strong barrier loci arise [8–10].

Recent empirical genomic studies have revealed regions of high and low differentiation dispersed throughout the genome at early stages of speciation [7,11,12]. This empirical data has lent strong support to the genic perspective of the speciation process [4]. To-date however, the majority of speciation genomic studies demonstrating heterogeneous genetic differentiation have come from young species or population pairs with low divergence [7,11,12]. Several thorough genomic studies on old sympatric species pairs exist, including European rabbits [13], *Drosophila* species [14], sunflowers [15], whitefishes [16], flycatchers [17,18], wild mice [19], Mimulus [20] and stick insects [9]; however except in a few cases, such as with Heliconius [21,22], divergence is thought to have occurred during periods of geographical isolation.

Distinction between primary and secondary divergence is important for interpreting the patterns of genomic differentiation [12,17]. This is because high genome-wide differentiation may have evolved via genetic drift and local adaptation during allopatric isolation, rather than due to divergence with gene flow. Following secondary contact after geographical isolation, heterogeneous genomic differentiation may arise due to introgression. Without a picture of the demographic history, this scenario may be indistinguishable from primary divergence [23]. Despite the fact that the expected pattern of genomic differentiation during speciation is influenced by the timing and duration of geographical isolation [7], testing different demographic histories has been somewhat neglected by the field [7,23], although this is now changing [17,24].

Other factors besides the demographic history of a species pair can also confound patterns of heterogeneous genomic differentiation. For example, variation in recombination rate

influences the patterns of genomic differentiation, because local adaptation or background selection in genomic regions where recombination is reduced can elevate differentiation measures and be mistaken for barrier loci [18,25,26]. Mutation rate variation also influences the patterns of absolute divergence [27]. Regions of low differentiation may be caused by shared ancestral polymorphism rather than gene flow [25,28]. Distinction between gene flow and shared ancestral polymorphism is likely easier in more divergent species pairs [27,29,30]. Furthermore, the use of multiple classical and recently developed methods, such as detection of recent hybrid progeny, ABBA-BABA tests [21,31], model-based inference [32], and comparisons between allopatric and sympatric pairs [21,26] provide a means to distinguish signatures of gene flow from alternative explanations. It is therefore essential to account for factors such as demographic history, recombination rate variation, and shared ancestral polymorphism that can confound the interpretation of genome scan data [7,12].

Three-spined stickleback species pairs (genus *Gasterosteus*) span the speciation continuum at varying stages of divergence, making them a model system for speciation research [33,34]. To-date genomic research on speciation with gene flow in the stickleback complex has largely focused on weakly divergent species pairs, such as lake-stream ecotypes [35–37]. Such studies have shown that the genomic landscape of differentiation between these recently diverged sympatric or parapatric species pairs is heterogeneous and interspersed with multiple peaks of high differentiation [35,37,38]. The emerging pattern is consistent with predictions under the genic concept of speciation–i.e. that reproductive isolation is localized in the genome at early stages of divergence [4,39]. However, it remains unclear whether such localized differentiation will eventually progress toward genome-wide differentiation in the face of gene flow [40].

Toward the end of the stickleback speciation continuum is a marine species pair in Japan [41,42]. The Japan Sea stickleback (*G. nipponicus*) is sympatric with the Pacific Ocean lineage of three-spined stickleback (*G. aculeatus*) (Fig 1A) in the waters surrounding the Japanese archipelago (Fig 1C) [41,43]. Divergence time between the two marine species has been estimated to be 1.5–2 million years based on allozyme and microsatellite data [42,44], making it much older than postglacial stickleback species pairs. Divergence between the species may have occurred as a result of the repeated isolation of the Sea of Japan during the Pleistocene, but this divergence scenario remains to be explicitly tested [42,44]. A unique feature of the *G. nipponicus* and *G. aculeatus* system, relative to postglacial stickleback species pairs, is that a neo-sex chromosome has arisen due to a fusion between a Y chromosome and a previously autosomal chromosome IX (chrIX) in the *G. nipponicus* lineage [41,45]. Furthermore, crosses between Japan Sea females and Pacific Ocean males show hybrid male sterility [42]. Previous quantitative trait locus (QTL) mapping identified QTL for courtship behaviour on the neo-X and hybrid male sterility on the ancestral-X. However, there are other isolating barriers, such as eco-geographical isolation, temporal isolation, and ecological selection against migrants [42,46,47]. The combination of these multiple barriers most likely contributes to the strong reproductive isolation in this system [41,48]. However, despite such strong divergence, hybrids have been observed where the two species co-occur in Northern Japan [41] and phylogenetic discordance between nuclear and mitochondrial loci suggests some history of introgression during speciation [49,50]. Although the Japanese species pair represents one of the furthest points of divergence within the stickleback species complex, speciation remains incomplete. The evolutionary history and genome-wide patterns of genetic differentiation and introgression of this strongly divergent species pair therefore remains an open question.

The aim of our study was to address this gap in our knowledge; i.e. to quantify the patterns of genomic differentiation and introgression at a later stage of the stickleback speciation continuum. To this end, we used previously published whole-genome sequences and newly acquired Restriction-site Associated DNA sequencing (RAD-seq) data from the Japanese

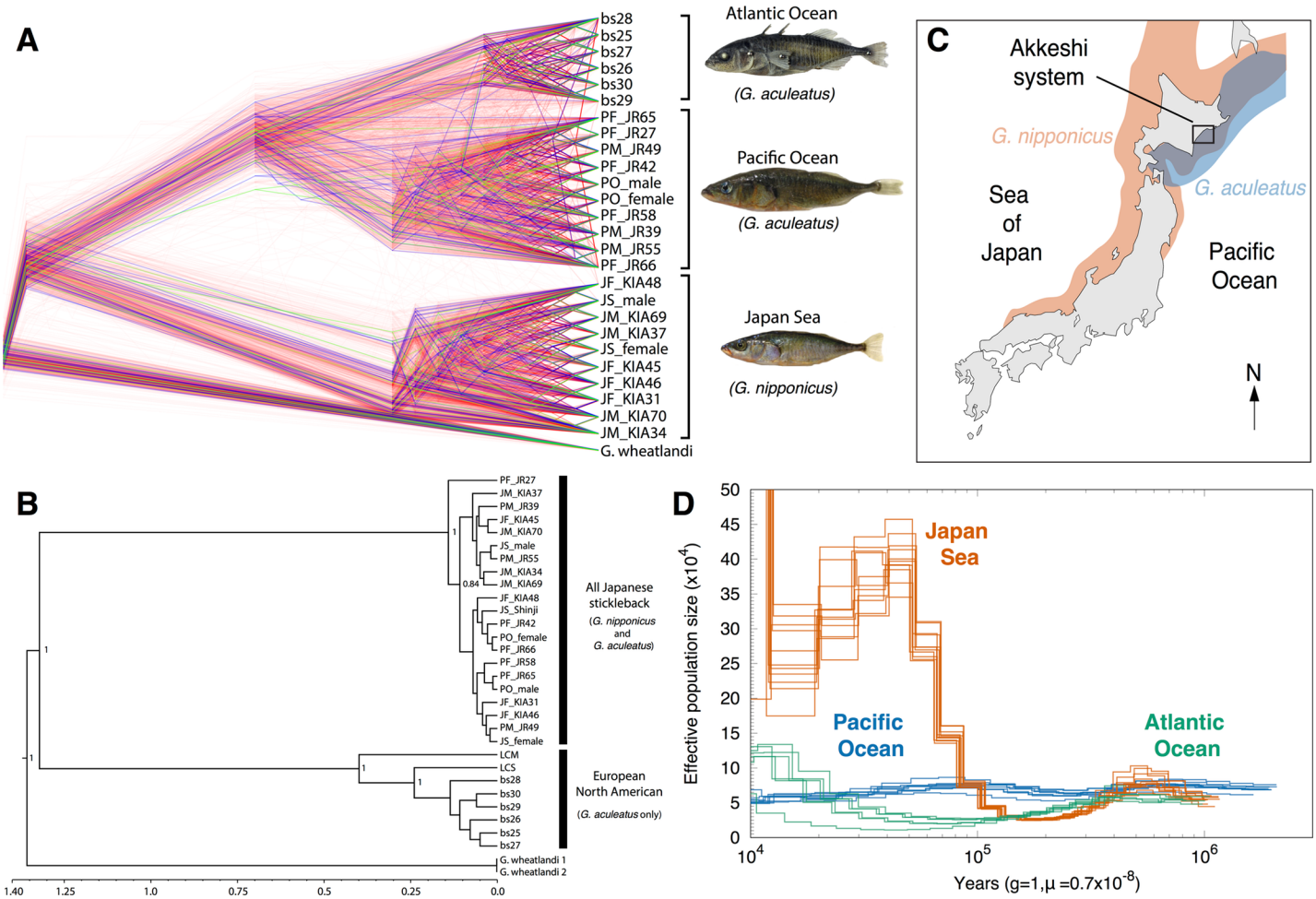


Fig 1. The Japan Sea stickleback is a separate species. (A) Rooted nuclear consensus tree for Japan Sea, Pacific Ocean and Atlantic Ocean stickleback lineages from 10 kb non-overlapping sliding windows across the autosomes. Red trees indicate species clustering; blue trees indicate geographical clustering and green trees reflect ancestral polymorphism. NB: Only 1,000 subsampled species trees are shown here to aid illustration. (B) Mitogenome Bayesian consensus tree shows divergence between two mitochondrial clades–all Japanese sticklebacks (*G. nipponicus* and *G. aculeatus*) and *G. aculeatus* occurring in Europe and North America. (C) Present day distribution of *G. aculeatus* (blue) and *G. nipponicus* (red) around the Japanese archipelago. The two species overlap in Hokkaido, Northern Japan and samples for this study were collected in Bekanbeushi River in Akkeshi unless noted. (D) PSMC plot of 26 resequenced genomes shows a steady effective population size in the Pacific Ocean lineage (blue) but a bottleneck around 0.15–0.3 million years before present and a subsequent increase in the Japan Sea lineage (orange). The effective population size of the Atlantic Ocean lineage is shown in blue green.

https://doi.org/10.1371/journal.pgen.1007358.g001

stickleback species pair to determine their evolutionary history and characterise patterns of gene flow between them. Our first aim was to establish how and when divergence took place between *G. nipponicus* and *G. aculeatus*. Using thousands of genomic loci and a coalescent modelling approach on the resequence data, we tested a range of divergence scenarios and estimated the timing and duration of isolation, the extent of gene flow and fluctuations in population size. After identifying that the two species have indeed diverged in the face of gene flow, we first used our RAD-seq dataset to investigate patterns of population structure and introgression between the Japanese stickleback species pairs. We then used a comparative genome scan approach with the resequence data, adding *G. aculeatus* lineage from the Atlantic Ocean [51] as an allopatric control (Fig 1A, S1 Fig). After establishing that gene flow has occurred but that a high level of genomic differentiation has remained, we used two independent measures of gene flow to identify where in the genome introgression has left its mark. We tested whether

introgression occurs more frequently in regions of high recombination and whether it occurs in regions with functionally important genes. Our findings suggest a high level of genome-wide divergence can be maintained in the face of gene flow, as introgression is restricted to small, localized genomic regions.

Results

Ancestral demography and population genomic analyses support divergence with gene flow

Phylogenetic analysis on 35,666 10 kb non-overlapping genome windows on autosomes (i.e., excluding chrIX and chrXIX) using whole genome resequence data on 26 individuals supports a deep split between *G. aculeatus* (both Pacific and Atlantic Ocean lineages) and *G. nipponicus* (Japan Sea stickleback) (Fig 1A). Of all windows, 98.8% support the split between species, while only 0.51% indicate clustering of fish occurring in Japan (the Japanese Pacific Ocean *G. aculeatus* and the Japan Sea *G. nipponicus*; S1 Table and Fig 1A).

We calculated genealogical sorting index (*gsi*) [52] on maximum likelihood phylogenies estimated from non-overlapping sliding windows of 10 kb across the autosomes. High *gsi* indicates monophyly, while low *gsi* indicates mixed ancestry [52]. Genome-wide averages (± SD) of *gsi* were high, but not complete, for all three *Gasterosteus* lineages with that of the Japan Sea stickleback being the highest (Atlantic *gsi* = 0.45 ± 0.10, Pacific *gsi* = 0.57 ± 0.09, Japan Sea *gsi* = 0.72 ± 0.06).

This is in stark contrast to the mitogenome phylogeny where sticklebacks from both species occurring in Japan fall into a single clade separate from the clade occurring in the Western Pacific and Atlantic (Fig 1B, S2 Fig). A lack of mitogenome divergence between *G. aculeatus* and *G. nipponicus* from the Japanese archipelago suggests mitochondrial introgression might occur where these lineages overlap (Fig 1C). Since the consensus autosomal phylogeny suggests a more recent split between the Japanese Pacific and Atlantic *G. aculeatus* lineages than the split in the mitochondrial phylogeny, the two mitogenome clades may represent the split between *G. aculeatus* and *G. nipponicus* lineages with mitochondrial introgression likely having occurred from the Japan Sea *G. nipponicus* into the Pacific Ocean *G. aculeatus* in sympatry. Divergence time estimates between the mitogenome clades are thus informative for dating the divergence time between *G. aculeatus* and *G. nipponicus* lineages. Bayesian coalescent analysis using a strict clock model in Bayesian Evolutionary Analysis by Sampling Trees (BEAST) suggests a median split date of 1.30 million years (0.15–2.41; 95% Highest Posterior Density [HPD] intervals; S2 Table) for the two major mitogenome clades (S2 Fig), consistent with previous estimates [49]. Divergence between Eastern Pacific and Atlantic haplotypes is more recent at 0.39 million years (0.03–0.74; 95% HPD) but is older than the Most Recent Common Ancestor (MRCA) of all haplotypes occurring in Japan (Fig 1B, S2 Fig), suggesting mitochondrial gene flow from *G. nipponicus* to *G. aculeatus* may have occurred within the last 0.39 million years.

To investigate the demographic history of *G. aculeatus* and *G. nipponicus*, we first used pairwise sequential Markov coalescent (PSMC) on all 26 Atlantic Ocean, Japan Sea and Pacific Ocean resequenced stickleback genomes to examine fluctuations in effective population size. Strikingly, *G. nipponicus* experienced a severe bottleneck around 0.15–0.3 million years before present (BP) (Fig 1D); mean N_e fell to 26,422 ± 1,191 at its lowest point. Subsequently after 0.1 million years BP, *G. nipponicus* underwent a dramatic effective population size expansion (Fig 1D): mean N_e rose to 195,974 ± 28,832 (i.e. ~7.5 times increase from the bottleneck) during the late Pleistocene. In contrast, the effective population size of the Japanese Pacific Ocean *G. aculeatus* has remained relatively stable throughout its history (mean N_e ± SD = 118,150 ± 4,330; Fig 1D, see S3 Fig for bootstrap support). Although the Atlantic (Fig 1D) and Western

Pacific lineages of *G. aculeatus* (S4 Fig) also experienced some growth during the late Pleistocene, their effective population sizes remained smaller than that of *G. nipponicus*. Cryptic population structure in *G. nipponicus* might explain the disparity in N_e between lineages; however our RAD-sequence dataset confirms substructure is not present in this species (see below for more details on RAD-seq dataset; S5 Fig and S6 Fig). Furthermore, genome-wide averages of Tajima's *D* also support a recent demographic expansion for *G. nipponicus* (mean ± SD of Tajima's *D* = -0.82±0.45) and stable effective population size in the Pacific Ocean (mean ± SD of Tajima's *D* = -0.04 ± 0.63).

To explicitly test whether divergence between *G. aculeatus* and *G. nipponicus* occurred in the presence of gene flow, we used an Approximate Bayesian Computation (ABC) approach with 1,874 2 kb loci randomly sampled from across autosomes. We tested five divergence scenarios–isolation (I), isolation with migration (IM), isolation-with-ancient-migration (IAM), isolation-with-recent-migration (IRM) and isolation-with-ancient-and-recent-migration (IARM)–i.e. two discrete periods of contact. Since the results of our PSMC analyses indicate N_e has varied throughout divergence (Fig 1D), we performed a hierarchical ABC analysis, first selecting the most appropriate population growth model (i.e. constant size, population growth and a Japan Sea bottleneck) within each divergence scenario and then performing final model selection amongst the best supported divergence/growth model scenarios (see S1 Text for full specification of models, priors, parameters and extensive sensitivity testing).

Using 20 summary statistics (see S1 Text for a full list of statistics used) and a neural-network rejection method with 1% tolerance of simulated datasets, the best-supported divergence scenario was a model of IM with a bottleneck occurring only in the Japan Sea species (Fig 2A, Table 1). An IARM model was the second best supported model. The use of a standard ABC rejection method gave rise to the qualitatively similar results and we found no evidence of an overrepresentation of introgressed regions in the loci used as the observed data for this analysis (S1 Text). An independent maximum likelihood based demographic analysis using the joint *G. aculeatus* and *G. nipponicus* site frequency spectrum (SFS) derived from RAD-seq data showed high support for an IARM model (see S1 Text).

Parameter estimates from the ABC IM model suggest divergence between *G. aculeatus* and *G. nipponicus* occurred 0.68 million years ago (median estimate, 0.18–4.17 million years, lower & upper 95% HPD; Fig 2B). A Japan Sea bottleneck occurred 0.3 million years ago (0.03–2.21 million years 95% HPD), reducing N_e to about 20% of the contemporary estimate (Fig 2C, S3 Table). Mean migration rates between the two species were low, and migration rate (expressed as m_{ij}–i.e proportion of population *i* that are migrants from *j* per generation) from the Pacific Ocean lineage into the Japan Sea lineage (m_{12}: median = 1.3×10^{-6}, 95% HPD = 8.61×10^{-8}–5.32×10^{-6}) was slightly greater than in the opposite direction (m_{21}: median = 1.05×10^{-6}, 95% HPD = 4.91×10^{-8}–6.39×10^{-6}, N.B. migration rates are backwards in time; see also Fig 2D & S3 Table). In addition to this, the distribution of the migration rate hyperprior suggested that a large number of loci showed some level of gene flow (S1 Text). Contemporary N_e of the Japan Sea lineage is larger than that of the Pacific Ocean, although the N_e estimates differed in magnitude from those estimated by PSMC (Figs 1D and 2C, S3 Table). Given this difference in effective population size, the scaled migration rates, the expected number of migrants per generation, ($2N_i m_{ij}$) are higher from Pacific Ocean lineage into the Japan Sea than the alternative (PO to JS = 0.18; JS to PO = 0.04) in contemporary populations, although still very low. Scaled migration rates were likely more similar during the Japan Sea bottleneck, because lower effective population size of the Japan Sea population (1.22×10^4) at this stage reduces the expected number of migrants from the Pacific Ocean to the Japan Sea (0.031).

Identifying admixture and the presence of backcrossed individuals between species where they co-occur provides strong evidence of on-going introgression [7,12]. To address this, we

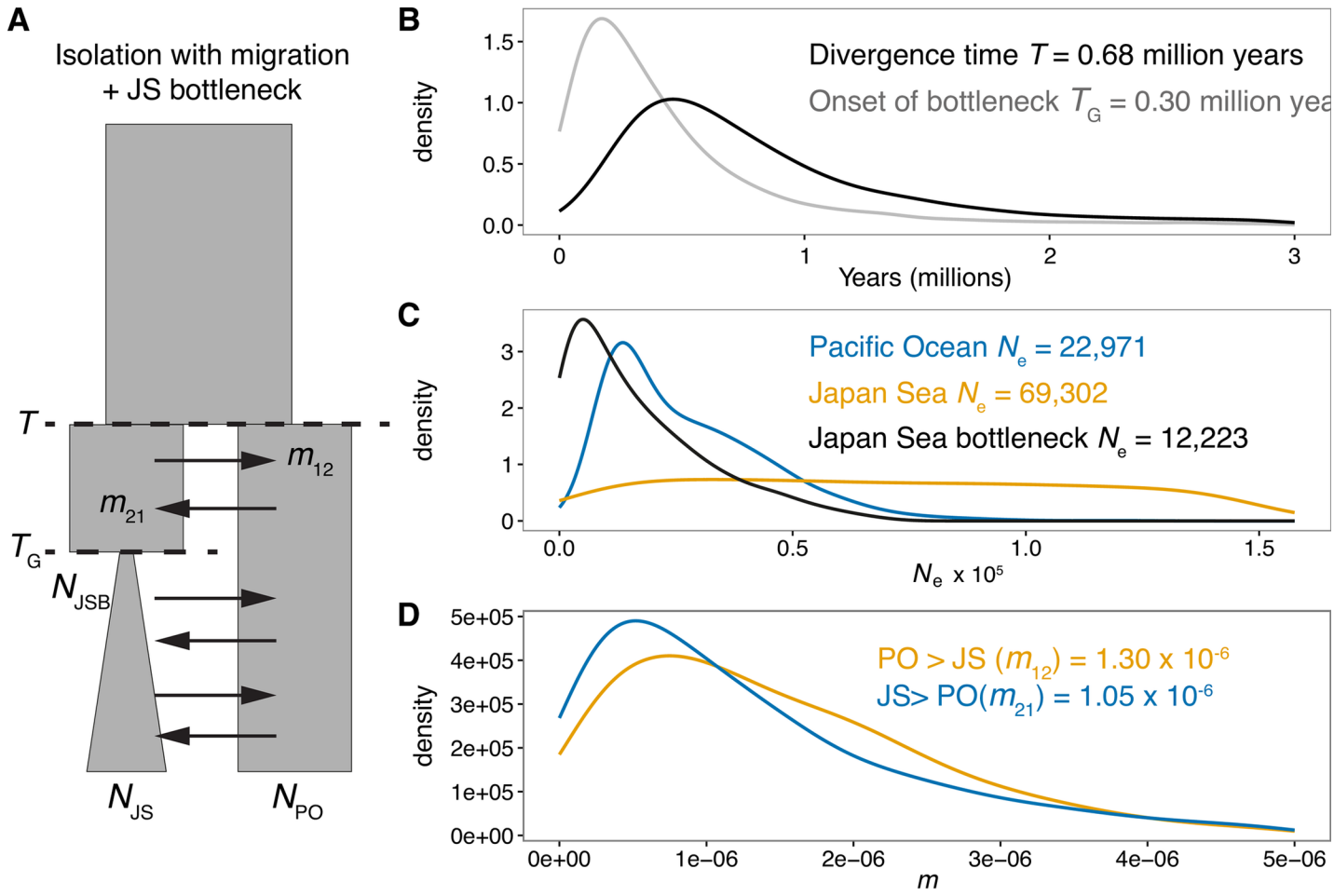


Fig 2. ABC analysis supports isolation with gene flow. (A) A model of isolation with migration and a bottleneck in the Japan Sea lineage is best supported by ABC analysis using ~2,000 nuclear loci (see Table 1). Posterior probability densities for model parameters estimated using neural network analysis with a tolerance of 1% and 20 summary statistics. Parameters are: T = time of split, m_{12} = the proportion of the Japan Sea population that are migrants from the Pacific Ocean per generation, m_{21} = the proportion the Pacific Ocean population that are migrants from the Japan Sea per generation (note that *m* is the migration rate backward in time); T_G = timing of bottleneck, N_{PO} = Pacific Ocean effective population size, N_{JS} = Japan Sea effective population size and N_{JSB} = Japan Sea bottleneck effective population size. Posterior probability density curves for (B) Japan Sea and Pacific Ocean divergence time and timing of bottleneck in the Japan Sea lineage, (C) Japan Sea, Pacific Ocean and Japan Sea bottleneck effective population sizes, and (D) migration rates averaged across the genome, shown as *m* in Fig 2A. Figures on each panel are median parameter estimates.

https://doi.org/10.1371/journal.pgen.1007358.g002

used a RAD-sequencing dataset with a larger sample size of 245 individuals from the Atlantic, Pacific and Japan Sea lineages, including previously published data from Pacific-derived populations in North America [53]. Principal component analysis (PCA) of allele frequencies at 3,

Table 1. Posterior probability values for models for final ABC model selection using neural network rejection. All estimates produced using a tolerance of 1% and 20 summary statistics. Bold text indicates the model where posterior probability provides the highest support. Models are I = isolation, IM = isolation with migration, IAM = isolation and ancient migration, IRM = isolation and recent migration, IARM = isolation with ancient and recent migration.

Divergence model	Growth model	Posterior probability
IM	bottleneck	**0.511**
I	bottleneck	0.008
IAM	bottleneck	0.009
IARM	bottleneck	0.343
IRM	bottleneck	0.129

https://doi.org/10.1371/journal.pgen.1007358.t001

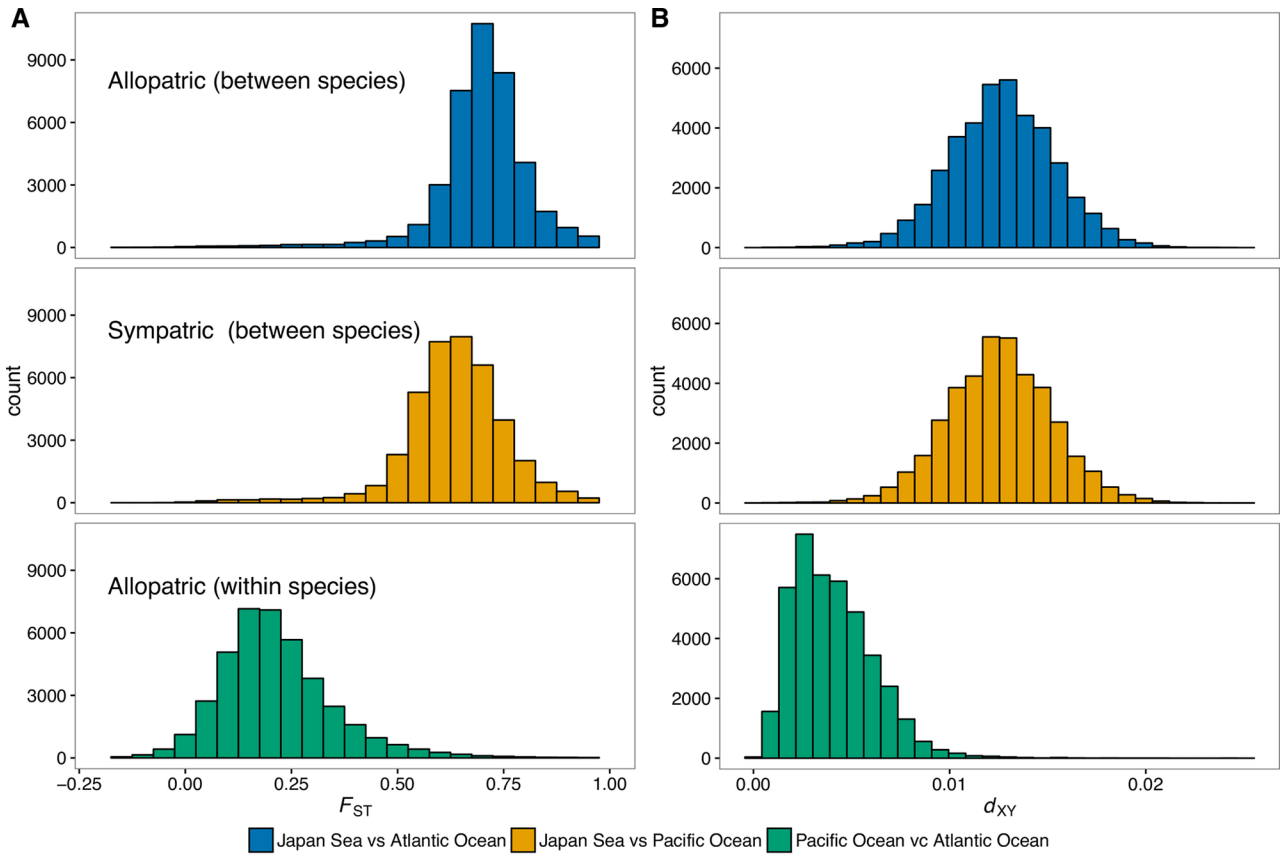
744 high-quality bi-allelic SNPs pruned to remove loci in linkage disequilibrium showed that, consistent with our whole genome data, the main axis explaining 20% of the variance was between *G. aculeatus* and *G. nipponicus* (S5 Fig). The secondary axis explaining 9.49% of the variance was mainly between the Atlantic and Pacific populations (S5 Fig). Importantly, PCA showed a single individual was intermediate between the Pacific and Japan Sea populations occurring in Akkeshi, the sympatric site in Hokkaido, Japan where our whole genome-sequenced samples were collected (Fig 1C). A separate Bayesian analysis for admixture using STRUCTURE [54,55] found greatest support for $K = 2$ among stickleback populations and also identified the putative F_1 hybrid plus individuals with possible recent admixture in Akkeshi (S6 Fig). To further investigate variation in individual ancestry, we identified 5,967 ancestry-informative loci i.e. autosomal SNPs with an allele frequency difference of >0.8 between the Japan Sea and Pacific Ocean lineages. Using a genomic cline approach, we estimated interspecific heterozygosity (i.e., proportion of loci with alleles from both species) and hybrid index (i.e., proportion of alleles from one species) on simulated hybrid genotypes. This indicates the marker set has high power to detect hybrid ancestry (S7 Fig). Analyses on the observed data suggest the RAD-seq dataset includes one F_1 hybrid and several individuals with likely hybrid ancestry in the last few generations (S7 Fig).

Taken together, these data indicate that divergence between the Japanese *G. aculeatus* and *G. nipponicus* is much older and greater compared to commonly studied postglacial stickleback species pairs. Despite the great extent of divergence between Japanese stickleback species, parameter estimates and observational data suggest that gene flow between them is on-going.

High levels of genome-wide divergence with highly localized signatures of introgression

Genome-wide differentiation was strikingly high between *G. nipponicus* and *G. aculeatus* regardless of their geographical overlap (Fig 3A & 3B and Fig 4, and S8 Fig and S9 Fig). The genome-wide average of F_{ST} between the sympatric species was 0.628; this is higher than F_{ST} in all other studied stickleback species pairs, which is typically less than 0.3 [35–37,56] (see Fig 3C). The genome-wide average of absolute divergence (d_{XY}) was 0.012; which is also high compared to previously calculated d_{XY} values, i.e. less than 0.005, between postglacial parapatric and sympatric stickleback ecotypes [35,57,58]. Despite consistently high divergence, both F_{ST} and d_{XY} values were significantly lower where the two species occur in contact (Table 2, Figs 3 &, 4, S8 Fig and S9 Fig; 10,000 replicate permutation tests on 10 kb windows: $P < 2.2 \times 10^{-16}$ for both statistics), consistent with the presence of gene flow in sympatry.

A more fine-scale analysis of genome-wide divergence based on 10 kb non-overlapping windows revealed that the high baseline divergence between *G. nipponicus* and *G. aculeatus* is interspersed by regions of low differentiation in both F_{ST} and d_{XY} genome scans (Fig 4 top two panels, S8 Fig and S9 Fig), possibly indicating introgression. To identify genomic regions of recent introgression, we calculated two independent measures. The first of these was G_{MIN}, the ratio of the minimum d_{XY} to the average d_{XY} [30]. Under strict isolation, minimum d_{XY} relates to the upper bound of divergence time between two populations, whereas when introgression occurs, minimum d_{XY} reflects the timing of the most recent migration event [30]. The second measure was f_d, an estimate of the proportion of introgressed sites in a genome window, calculated using a four population ABBA-BABA test [59]. G_{MIN} is more effective at identifying recent, low level gene flow than either F_{ST} or d_{XY} but by definition it is unable to detect genomic regions where complete introgression has occurred [30], which can however be detected using f_d. Importantly, both measures are robust to variation in recombination rate [30,59]. Combining these two statistics therefore allows us to identify both low-level (G_{MIN}) and strong introgression (f_d).


A
B
Allopatric (between species)
Sympatric (between species)
Allopatric (within species)
count
F_{ST}
d_{XY}
Japan Sea vs Atlantic Ocean
Japan Sea vs Pacific Ocean
Pacific Ocean vc Atlantic Ocean

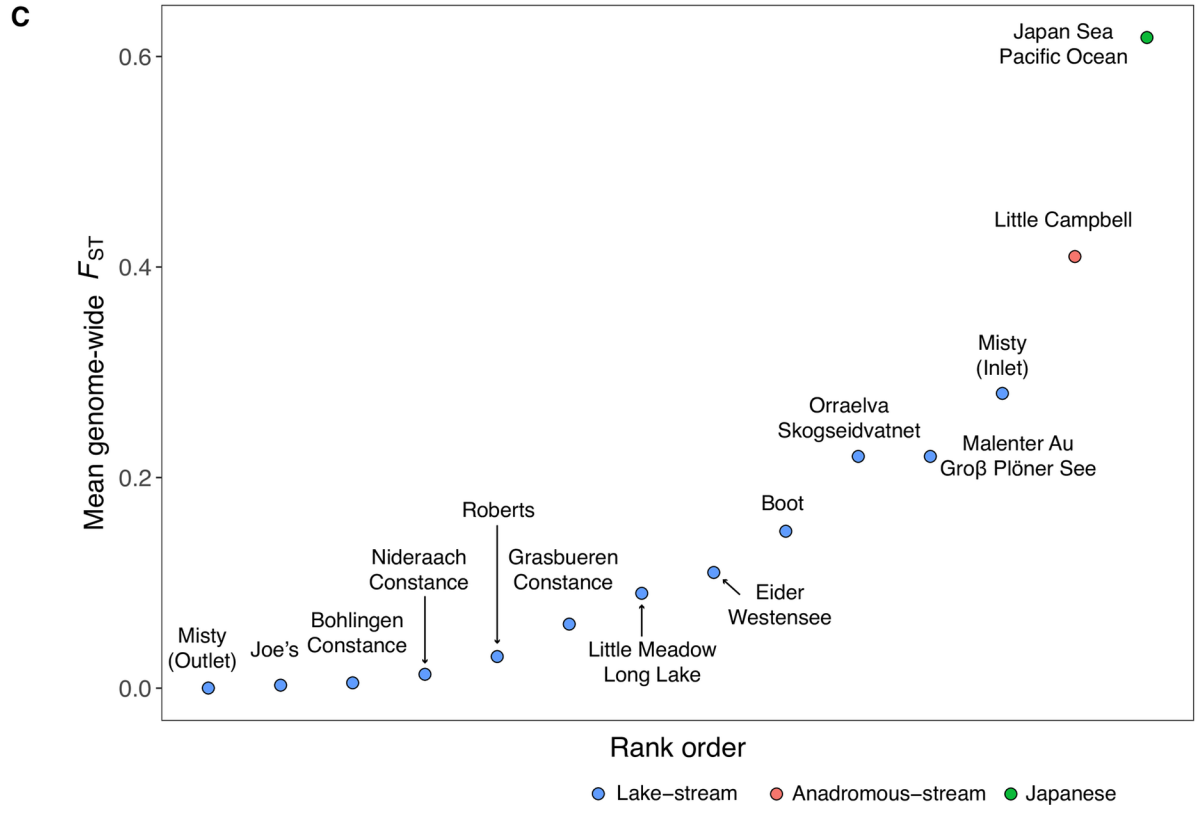

C
Mean genome-wide F_{ST}
Rank order
Japan Sea
Pacific Ocean
Little Campbell
Misty
(Inlet)
Orraelva
Skogseidvatnet
Malenter Au
Groβ Plöner See
Boot
Eider
Westensee
Roberts
Nideraach
Constance
Grasbueren
Constance
Bohlingen
Constance
Misty
(Outlet)
Joe's
Little Meadow
Long Lake
Lake–stream
Anadromous–stream
Japanese

Fig 3. Genomic divergence is lower in sympatry than in allopatry between species. Histograms of (A) relative (F_{ST}) and (B) absolute (d_{XY}) differentiation measures for each of the species comparisons. (C) Mean genome-wide F_{ST} of the Japanese species pair compared with those of other stickleback systems taken from previously published studies [35–37,56].

https://doi.org/10.1371/journal.pgen.1007358.g003

Focusing on between species comparisons, mean (± SD) G_{MIN} measured from 10 kb non-overlapping windows was greater in allopatry than sympatry (Japan Sea vs. Atlantic: 0.876 ± 0.071; Japan Sea vs. Pacific: 0.857±0.103; randomization test $P < 2.2x10^{-16}$; Fig 4). Mean f_d was also greater when the species overlapped (JS vs. AT: -0.0031 ± 0.0540; JS vs. PO: 0.0039 ±0.0328; $P < 2.2x10^{-16}$; Fig 4), and both statistics are more strongly negatively correlated in sympatry (S10 Fig) supporting gene flow between *G. nipponicus* and Japanese populations of *G. aculeatus*.

Genomic regions of low G_{MIN} (i.e. G_{MIN} valleys) may indicate recent introgression. We identified genome windows with low G_{MIN} values using a Hidden-Markov classification model [60] (S11 Fig). We then clustered 10 kb outlier windows occurring within 30 kb of one another into putative G_{MIN} valleys. G_{MIN} in particular may be susceptible to false positives as a result of shared ancestral polymorphism. However, lower d_{XY} and higher f_d in sympatric G_{MIN} valley windows compared to the genomic background suggests shared ancestral polymorphism alone does not explain the patterns observed here (S12 Fig; randomization test, $P < 2.2 \times 10^{-16}$ in both cases). These regions of introgression were more common in the genome when the two species overlapped, with 637 valleys in sympatry (JS-PO comparison) compared to 337 in allopatry (JS-AT comparison) (randomization test, t = 5.35, $P < 2.2 \times 10^{-16}$) and a greater number of valleys per chromosome (Fig 5A), although mean valley size did not differ significantly (77.6 kb and 75.4 kb in sympatry and allopatry respectively, $P = 0.82$). Interestingly, 225 valleys were shared between JS-PO and JS-AT comparisons (Fig 4). These shared valleys may indicate shared ancestral polymorphism but they may also reflect introgression from Pacific Ocean to Japan Sea, where one or a few Japan Sea individuals carry haplotypes derived from Pacific Ocean and therefore are also similar to Atlantic Ocean haplotypes too. However, a larger number of valleys (412 valleys) were unique to the JS-PO comparison, where introgression might occur from Japan Sea to Pacific Ocean.

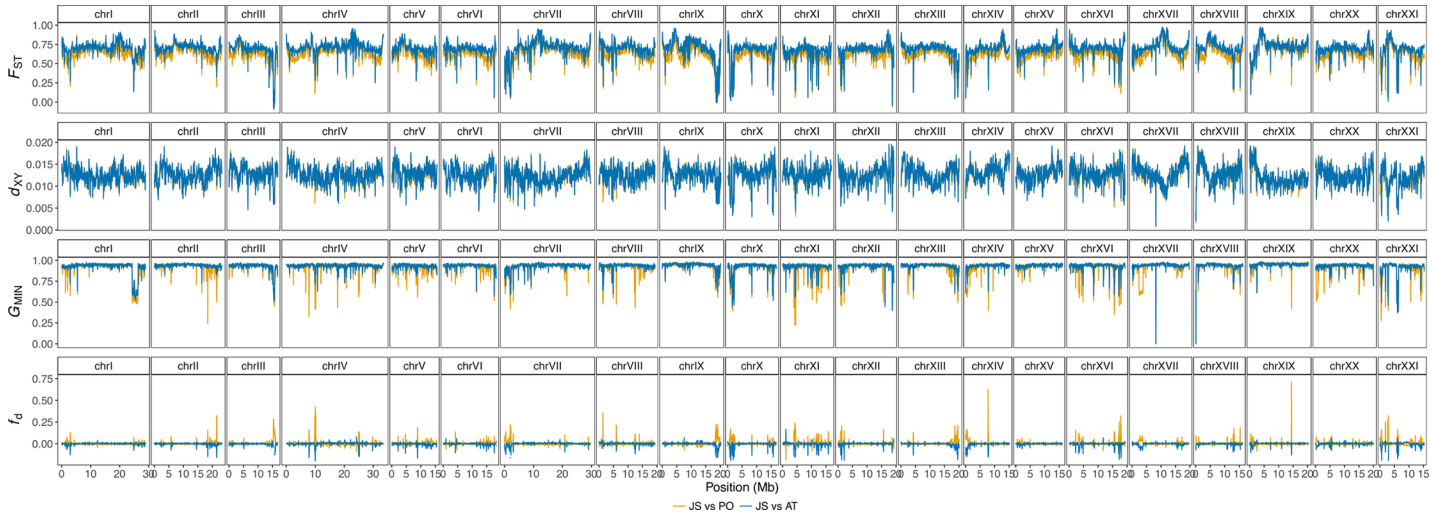


Fig 4. Genome-wide distribution of divergence and introgression. Divergence was measured using F_{ST} and d_{XY}, while introgression was measured using G_{MIN} and f_d. Data plotted here is from 50 kb non-overlapping genome windows. Blue and yellow lines indicates allopatric (Japan Sea vs Atlantic) and sympatric (Japan Sea vs Pacific Ocean) comparisons, respectively.

https://doi.org/10.1371/journal.pgen.1007358.g004

Table 2. Genome-wide averages for measures of divergence and introgression. F_{ST}, d_{XY}, G_{MIN}, and f_d for all pairwise comparisons of Japan Sea (JS), Pacific Ocean (PO) and Atlantic Ocean sticklebacks (AT) are shown. Mean ± SD and lower and upper limits of the 95% confidence interval (in parenthesis) are shown. NA, not analysed.

Comparison	F_{ST}	d_{XY}	G_{MIN}	f_d
JS vs PO	0.634 ± 0.122 (0.333–0.862)	0.012 ± 0.002 (0.007–0.017)	0.857 ± 0.102 (0.513–0.942)	0.004 ± 0.054 (-0.031–0.085)
JS vs AT	0.697 ± 0.116 (0.406–0.902)	0.013 ± 0.002 (0.007–0.018)	0.876 ± 0.071 (0.666–0.942)	-0.003±0.033 (-0.077–0.029)
PO vs AT	0.215 ± 0.134 (0.003–0.539)	0.004 ± 0.002 (0.001–0.009)	0.560 ± 0.141 (0.223–0.772)	NA

https://doi.org/10.1371/journal.pgen.1007358.t002

A similar geographical comparison of peaks of f_d between species was not possible, due to the fact that f_d is much closer to 0 in the comparison between *G. nipponicus* and the Atlantic *G. aculeatus* and very few peaks were present (Fig 4). Nonetheless, Hidden-Markov classification identified 823 f_d peaks occurring between *G. nipponicus* and Pacific *G. aculeatus* (S13 Fig). If the f_d peaks mainly indicate introgression from Pacific Ocean to Japan Sea, d_{XY} between Japan Sea and Atlantic Ocean is expected to be lower in these regions compared to the genome background, as Japan Sea fish carry haplotypes derived from the Pacific Ocean, which in turn are similar to the Atlantic Ocean haplotypes. While JS-AT d_{XY} was lower in f_d peaks compared to the genome background (JS-AT mean d_{XY} ± SD, f_d peaks: 0.0121±0.0035, genome-background: 0.0127±0.0026; one-tailed permutation test, $P < 2.2 \times 10^{-16}$), this difference was not very clear (S14 Fig). In contrast, if introgression occurred mainly from Japan Sea to Pacific Ocean, d_{XY} in the PO-AT comparison should increase in f_d peaks relative to the genome background, as Pacific Ocean fish carry Japan Sea-derived haplotypes, which are divergent from

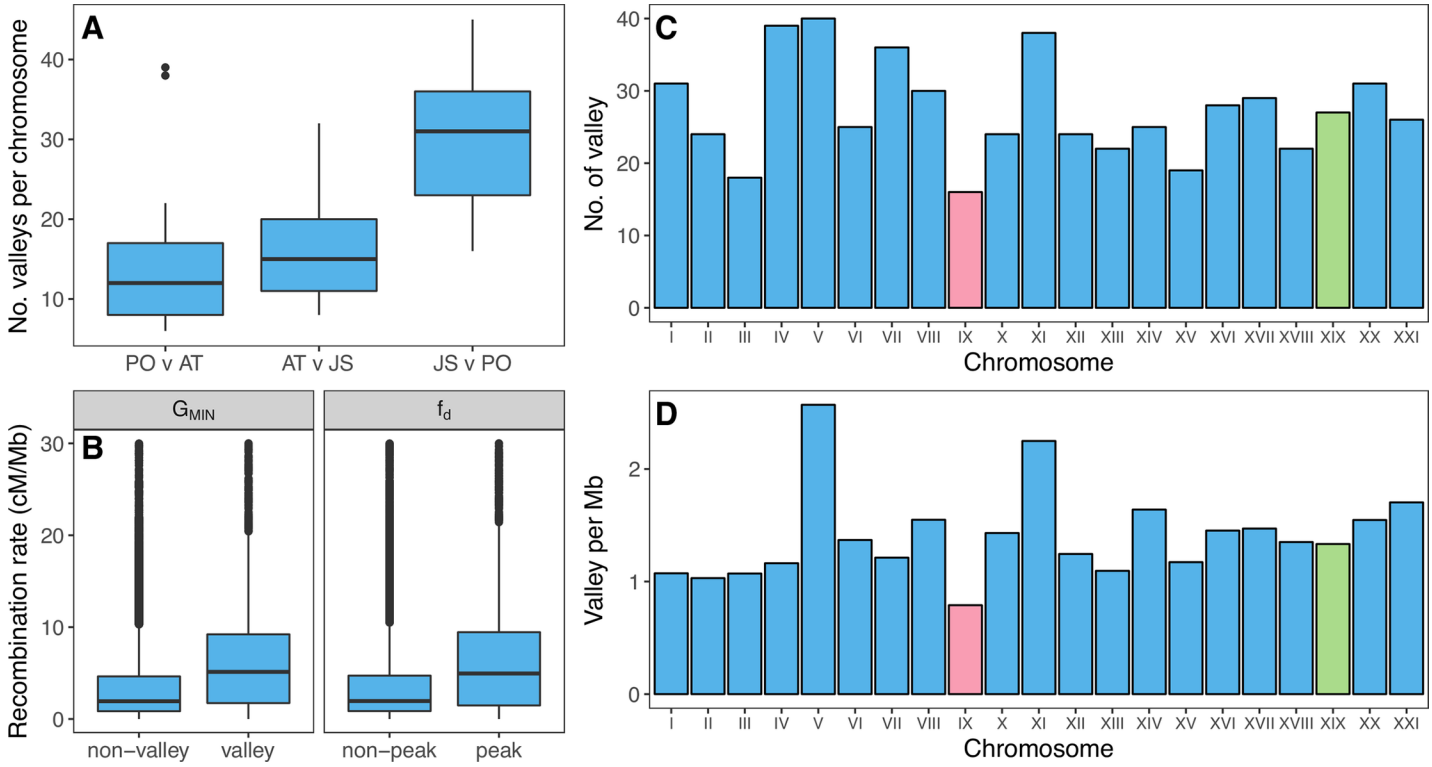


Fig 5. Fewer introgression valleys occur on the neo-X chromosome. (A) A greater number of G_{MIN} valleys occur in sympatry than in allopatry between species. (B) G_{MIN} valleys and f_d peaks also occur in regions of the genome with a higher recombination rate. Fewer valleys occur on the neo-X chromosome (chrIX; shown in pink) compared to autosomes (C), even when chromosome length is taken into consideration (D); N.B.–data for (C) and (D) were measured using females only. Green shows the ancestral sex chromosome (chrXIX).

https://doi.org/10.1371/journal.pgen.1007358.g005

the Atlantic Ocean haplotypes. We clearly observed this pattern (PO-AT mean d_{XY} ± SD, f_d peaks: 0.0065±0.0035, genome-background: 0.0038±0.00182; $P < 2.2 \times 10^{-16}$; S14 Fig); suggesting that introgression from Japan Sea to Pacific Ocean may be more predominant than the opposite direction. Importantly, our findings using G_{MIN}, d_{XY} and f_d are robust to different missing data thresholds and did not change when phased vs. unphased data is used (S1 Text).

To further investigate the direction of gene flow, we used partitioned *D* statistics (an extension of the four population test–see S15 Fig), which tests the excess of shared derived alleles using five, rather than four populations [61]. To this end, we added an allopatric Japan Sea population (collected from Lake Shinji, a brackish lake at the Japan Sea coast of southern Honshu). A positive D_{12} statistic is proposed to indicate the predominance of introgression from P3 to P2 (S15 Fig) [61]. When P3 was set to Japan Sea (where $P3_1$ is sympatric and $P3_2$ is allopatric with the Pacific Ocean) and P2 to Pacific Ocean (see S14 Fig), D_{12} was significantly positive in f_d peaks (one-tailed permutation test, $P < 2.2 \times 10^{-16}$). In contrast, when we rotated the populations at the tips–i.e. setting P2 to sympatric Japan Sea, $P3_1$ to Pacific Ocean, and $P3_2$ to Atlantic Ocean (see S15 Fig), D_{12} was not positive, consistent with the suggestion that introgression is occurring mainly from Japan Sea to Pacific Ocean. However, the resolution of partitioned *D* statistics has been criticized [62]; positive D_{12} can also be caused by introgression from the Pacific Ocean (P2) to the common ancestor of the sympatric and allopatric Japan Sea populations ($P3_1$ & $P3_2$). To overcome this issue, we calculated D_{FOIL}, which also uses a five-population test but accounts for all possible introgression events [62]. When P_1 = sympatric Japan Sea, P_2 = allopatric Japan Sea, P_3 = Pacific Ocean, and P_4 = Atlantic Ocean (S16 Fig), D_{FOIL} clearly indicated the presence of ancestral introgression (239 out of 4,236 100 kb-windows) between the Japan Sea ancestor (P_{12}) and the Pacific Ocean (P_3) (see S15 Fig). However, we found only a few windows showing unidirectional introgression (6 in total), meaning we cannot determine the direction of introgression using this analysis (S16 Fig). This low sensitivity may be due to the fact that structuring in the Japan Sea lineage is very low (S6 Fig) [63]–i.e. recent divergence time between the sympatric and allopatric Japan Sea populations or high intraspecific gene flow within the Japan Sea species.

Characterization of genomic regions of introgression

To investigate whether introgression co-varies with recombination rate, we used a previously published recombination map from an Atlantic *G. aculeatus* cross [64] to interpolate genome-wide recombination rate variation (see Methods). We detected a negative correlation between recombination rate and G_{MIN} and a positive correlation with f_d (Pearson's correlation, G_{MIN}: r = -0.17, $P < 2.2 \times 10^{-16}$; f_d: r = 0.08, $P < 2.2 \times 10^{-16}$, S17 Fig). Accordingly, mean recombination rate for putatively introgressed regions was over two times higher than the genome background (G_{MIN}: valley = 8.98 cM/Mb, non-valley = 3.99 cM/Mb; f_d: peak = 9.64 cM/Mb, non-peak = 4.16 cM/Mb; randomization test $P < 2.2 \times 10^{-16}$ in both cases; Fig 5B).

Sex chromosomes likely played an important role in speciation between *G. aculeatus* and *G. nipponicus* [41,45]. A fusion between Y and chrIX means that chrIX segregates as a neo-sex chromosome in *G. nipponicus* but not *G. aculeatus* which only carries the ancestral and shared sex chromosome, chrXIX [41,45]. The divergent XY (*G. aculeatus*) and X_1X_2Y (*G. nipponicus*) systems means that recombination is reduced for chrIX and chrXIX in hybrids carrying the neo-Y [45]. Given this recombination rate reduction and previously identified QTL for traits involved in reproductive isolation that map to chrIX and chrXIX [41,45], we tested whether recent introgression (i.e. measured using G_{MIN}) was reduced in this part of the genome relative to the autosome. For this, we repeated our analyses using females only (5 Japan Sea and 6 Pacific Ocean). The number and density of valleys was lowest on the neo-sex chromosome,

chrIX (16 valleys or 0.8 valleys per Mb) but not on the ancestral sex chromosome (chrXIX, see S4 Table).

Finally, we investigated the nature of introgression between the two species. We first asked whether introgression occurs more frequently in genic or non-genic regions. We identified 3,261 genes occurring in G_{MIN} valleys and 2,958 genes from f_d peaks between sympatric *G. aculeatus* and *G. nipponicus*; 60% of genes identified were found in both types of introgressed window, whereas 23% occurred only in G_{MIN} valleys and 15% only in f_d peaks (S18 Fig). Irrespective of the method used to detect putatively introgressed regions, the number of genes identified was greater than the number expected by chance ($P < 0.0001$ based on a null distribution generated from 1,000 random samples of the genome). Mean recombination rate was higher in the genomic windows where genes are present compared to the genomic background (gene windows = 4.92 cM/Mb, genome-background = 4.24 cM/Mb; permutation test: $P < 2.2 \times 10^{-16}$). This suggests that introgression may be more likely in genic regions of the genome than non-genic regions, which can be partly explained by higher recombination rates in genic regions.

To further investigate the functional enrichment of the genes occurring in regions of introgression, we performed gene ontology (GO) analysis on 2,310 G_{MIN} valley and 2,217 f_d peak genes with orthologs in the human genome. Enriched GO terms for f_d peaks included immune response, metabolic processes and chromatin assembly, while enriched GO terms for G_{MIN} valleys included major histocompatibility complex (MHC) protein and metabolic processes (S5 Table & S6 Table).

Discussion

Japanese stickleback speciation has occurred in the face of on-going gene flow

Determining the demographic and evolutionary history of species pairs is an important first step for understanding how speciation has unfolded in any system [7,12]. Our present study has produced several lines of evidence indicating that divergence between the Japanese sticklebacks has occurred in the presence of gene flow.

Firstly, our ABC analysis supported a model of isolation with migration. Previously, it has been speculated that the Japan Sea stickleback diverged largely as a result of geographical isolation in the Sea of Japan caused by sea level fluctuation during the early Pleistocene [42,44]. Using ABC, we were able to explicitly test several divergence hypotheses in a statistical framework [65]; our findings suggest that gene flow has likely occurred throughout majority of the divergence history. It should be noted that ABC and most established demographic inference methods perform poorly when resolving the timing of gene flow between lineages [66,67]. Therefore, one caveat to the interpretation of our ABC results is that we cannot rule out the possibility that the two species diverged in repeated cycles of contact (i.e. akin to our IARM model which had the second highest level of support; Table 1), but these periods of contact were simply too close in time. Our independent SFS-based demographic analysis using RAD-seq data also suggested higher support for an IARM model than for an IM model. Nonetheless, the posterior probabilities from models with migration in the ABC analysis overwhelmingly support a scenario of divergence with a period of gene flow irrespective of the timing or nature of the actual speciation event.

The presence of extant recent hybrids in sympatry also strongly indicates that introgression is still on-going. In several cases of sympatric pairs of highly diverged species [68–70], hybrids beyond F_1 are found and provide strong evidence for on-going gene flow. We observed a probable F_1 hybrid in the wild and several other individuals with evidence of recent hybrid ancestry

in our RAD-seq dataset, consistent with previous studies that observed wild caught hybrids [41,71]. This provides direct observation of admixture in the wild.

Lower levels of genome-wide divergence (both F_{ST} and d_{XY}) between sympatric pairs compared to allopatric pairs also indicate the presence of gene flow. Our G_{MIN} and f_d genome scans showed a higher number of putatively introgressed regions between *G. nipponicus* and Japanese Pacific *G. aculeatus* than between *G. nipponicus* and Atlantic *G. aculeatus*, suggesting that introgression has been occurring even after the Atlantic and Pacific stickleback populations diverged approximately 390,000 years BP. Our partitioned *D* statistics demonstrated that gene flow from *G. nipponicus* into Japanese Pacific *G. aculeatus* may be more predominant than the opposite direction in sympatry.

Contrasting mitochondrial and nuclear genome phylogenies are also consistent with the presence of gene flow. Mitochondrial introgression has likely occurred from *G. nipponicus* into *G. aculeatus* at some point in the last 0.39 million years. Our mitogenome phylogeny confirmed previous findings that there is no mitochondrial structure that distinguishes between the *G. nipponicus* and Japanese populations of *G. aculeatus* [49,50]. This is in contrast to our nuclear autosomal phylogeny which showed that majority of the genome supports a clear split between *G. nipponicus* and *G. aculeatus* occurring in Japan and that the latter shares a more recent common ancestor with Atlantic European *G. aculeatus* populations. In short, mitogenome data clusters the *Gasterosteus* lineages by geography, while the nuclear data clusters them by species. Disparities in effective population size between lineages are a common cause of unidirectional mitonuclear introgression with introgression likely occurring from a larger to a smaller population [72]. Our reconstruction of temporal variation in effective population size using PSMC showed a rapid population expansion of *G. nipponicus* during the late Pleistocene that created a large demographic disparity with the *G. aculeatus* Pacific Ocean lineage. Although it should be noted that admixture and cryptic population structure can increase effective population size estimates when using PSMC [73,74], we found no evidence of clear population structure in the Japan Sea lineage (S6 Fig). Furthermore, both Japan Sea and Pacific Ocean individuals were sequenced to very high mean coverage (80X), therefore differences in depth of coverage are very unlikely to explain the PSMC results [75] or introduce bias into our ABC analysis. Unidirectional mitochondrial introgression might also be caused by female mate choice [76]. Our previous behavioural studies indicate that Japan Sea females often mate with Pacific Ocean males, while Pacific Ocean females rarely mate with Japan Sea males [41,42]. Hybrid females from Japan Sea female and Pacific Ocean male crosses are fertile [42] and will carry Japan Sea mitochondrial DNA. Backcrossing of these hybrids to Pacific Ocean males would result in unidirectional mitochondrial introgression from the Japan Sea to Pacific Ocean.

High genomic divergence at a late stage of speciation with gene flow

Compared to young species pairs, less is known about the patterns of genomic differentiation at more advanced stages of speciation with gene flow. Our ABC analyses placed the estimated divergence time of *G. aculeatus* and *G. nipponicus* at 0.68 million years BP. Similarly, our Bayesian coalescent analysis of mitogenome divergence revealed a 1.3 million year split between the Japanese and Atlantic-Pacific *Gasterosteus* mitochondrial clades. Both mitochondrial and nuclear split estimates suggest that divergence between *G. aculeatus* and *G. nipponicus* occurred well before the end of the last glacial period. Therefore the Japanese stickleback system is older than all other previously examined postglacial sympatric or parapatric species pairs, which have typically diverged within the last 20,000 years [33].

The Japanese stickleback system also has a mean genome-wide F_{ST} and d_{XY} values higher than any other sympatric or parapatric stickleback species pair studied so far such as lake-

stream or freshwater-anadromous pairs (Fig 3C) [36,38,57]; placing this pair at the furthest end of the speciation continuum. The primary explanation for the observed elevated divergence is most likely the more ancient divergence time of the Japan Sea-Pacific Ocean species pair compared to postglacial species pairs [38,77]. However, the results of our demographic analyses indicate that high divergence is not due to a long period of allopatric isolation without gene flow, contrary to what has previously been suggested [42,44]. This is important, as failing to account for variation in evolutionary history among species pairs placed on a continuum will obscure the processes leading to higher differentiation as speciation progresses. A further explanation for the high genomic divergence is the presence of strong isolating barriers between the Japan Sea and Pacific Ocean sticklebacks. Total reproductive isolation (0.970) is greater than in all postglacial species pairs (0.716–0.895) [48] and arises from a combination of habitat [46,47], temporal [78] and sexual isolation, and hybrid sterility [41,42]. Recent theoretical studies have shown that selection on many barrier loci in the face of gene flow may result in a transition from low to high differentiation as a result of 'genome-wide congealing' [10,79]. It is important to note however that we lack evidence that such a transition might explain the high differentiation we see here relative to the rest of the stickleback continuum (Fig 3C).

Localized introgression at a late stage of speciation with gene flow

Our study has also demonstrated two important signatures of introgression in the Japanese sympatric stickleback pair. Firstly, levels of background genome differentiation between *G. aculeatus* and *G. nipponicus* estimated by F_{ST} were lower in sympatry compared to allopatry. We note that this pattern was observed both in our whole genome and RAD-seq datasets. The higher overall genetic differentiation between *G. nipponicus* and Atlantic *G. aculeatus* is likely due to genetic drift and local adaptation and the fact that these two lineages have never overlapped geographically. Secondly and strikingly, using resequencing data, we identified small regions of localised introgression dispersed throughout the genome when *G. nipponicus* and *G. aculeatus* co-occur in sympatry. These introgression regions were measured using $G_{MIN,}$ the ratio of minimum d_{XY} to mean d_{XY} [30], and $f_{d,}$ the proportion of introgressed sites in a genome window [59].

Several methodological issues might influence these measures of introgression. Firstly, there is a coverage disparity between resequenced individuals sampled in Japan and those from the Atlantic (mean 61X and 12X coverage respectively), but both sample sets are sequenced to a depth suitable for accurate genotyping. Furthermore, Atlantic Ocean individuals with relatively lower depth are not included in the analysis of ABC and only serve as a comparison for genome-wide patterns of differentiation, divergence or introgression between the sympatric Japanese species. Secondly, both G_{MIN} and f_d are sensitive to sample size; fewer individuals will mean rare haplotypes have a lower sampling probability. However, by re-conducting our analyses using only females, a much smaller sample size than our main analysis, we still identified clear signals of introgression. Thirdly, G_{MIN} will be biased downwards if a recently backcrossed individual is included in the dataset. All Japanese *G. aculeatus* and *G. nipponicus* used in the study were identified as 'pure' individuals with genotyping at multiple microsatellite loci prior to resequencing [41,45]. To further ensure that a single backcrossed individual was not biasing our findings, we examined the two haplotypes producing the lowest value of d_{XY} in each G_{MIN} valley to confirm that the majority were not always from the same individuals (doi:10.5061/dryad.104g3d0). Finally, shared ancestral polymorphism cannot explain why more G_{MIN} valleys occur in sympatry (Fig 5A) (S9 Fig & S13 Fig).

What then underlies the localised pattern of introgression we observe? One possible explanation is the fact that many isolating barriers are involved in reproductive isolation [41,48].

Although the genomic basis of these isolating barriers remains unknown, it is likely that barrier loci occur throughout the genome; pervasive selection at multiple loci is expected to limit the extent of introgression at this scale [80]. We found significant positive relationships between recombination rates and introgression. The strength and extent of negative selection against an allele at a barrier locus and genomic regions linked to it is inversely proportional to recombination rate [80]. Recombination determines effective migration rate [81]; when recombination is high, neutral and adaptive loci linked to the target of negative selection in the recipient population have a greater probability of escaping removal and so their probability of introgression is greater [3]. Selection has a higher efficiency in these high recombination rate regions due to increased effective population size–therefore deleterious introgression is also more likely to be removed. The expectation then is that signatures of introgressed neutral or adaptive alleles are most likely to persist in regions of the genome where recombination rate is sufficiently high enough, and indeed, the positive association between introgressed regions and recombination rate we observed supports this (Fig 5B, S15 Fig). Introgression is typically lower on sex chromosomes relative to autosomes in multiple taxa due to the effects of reduced recombination and greater exposure to selection in the hemizygous sex [82]. The sex chromosomes play an important role in the Japanese stickleback system, harbouring QTL for hybrid sterility and behavioural isolation [41]. Consistent with this, we observed lower introgression on the neo-sex chromosome (Fig 5E & 5F), although we cannot exclude the possibility that the fusion occurred more recently than the speciation event, so the opportunity for introgression on the neo-sex chromosomes was simply low relative to the rest of the genome. Taken together, our findings suggest that strong divergent selection and recombination rate variation may determine the localised signature of introgression in the genome.

The nature of gene flow in the Japanese stickleback system may also give some clues as to why we observe such highly localised introgression. One possibility is that a proportion of the introgression we detected is adaptive; i.e. it is maintained because of either directional or balancing selection. Adaptive introgression has been detected in a wide range of taxa [83], including humans [84]. However, the expected signatures of the process remain unclear–especially when introgression is widespread in the genome, as is possibly the case here. Our GO analyses suggest an enrichment of immune response genes, including MHC genes, and metabolism genes in introgressed regions. Immune genes have been identified as being under balancing selection in hybridising taxa, particularly plants [85] and birds [86]. Several genes involved in metabolism are also reported to be under balancing selection in humans [87]. Furthermore, recent analysis suggests that negative frequency dependent selection might result in introgression of rare MHC alleles between divergent stickleback ecotypes [88]. Further research is necessary to directly test whether this process might explain introgression in the Japanese stickleback system.

Conclusion

Much of our knowledge of how genomic differentiation builds along the speciation continuum is drawn from studies focusing on young, allopatric or completely reproductively isolated species pairs. Very few examples of species pairs at a later stage of divergence with on-going gene flow have been investigated. Here, we have shown that the Japan Sea and Pacific Ocean species pair exemplifies this under-represented stage of speciation and is situated at the further end of the stickleback species continuum. The high genomic differentiation between the species may be due to a more ancient divergence time than previously studied postglacial species pairs, selection on multiple isolating barriers or a combination of the two. Despite high differentiation, gene flow is on-going between the species and we identified localized signatures of

introgression throughout the genome. Although the localized nature of the introgression remains unclear, selection–either directional or balancing–may play some role in promoting it. Overall, our study demonstrates that high levels of genomic divergence can be established and maintained in the presence of gene flow. Further genomic studies on more species pairs at late stages of speciation with gene flow will help to understand the generality of the patterns seen here.

Materials and methods

Ethics statement

All animal experiments were approved by the institutional animal care and use committee of the National Institute of Genetics (23–15, 24–15, 25–18).

Sample collection, whole genome resequencing and RAD sequencing. Collection and sequencing of all Japanese individuals used for whole genome resequencing has been described previously [45] except the allopatric Japan Sea fish. Briefly sympatric populations were captured from the Akkeshi system in Hokkaido, Japan in 2006 (Fig 1C). The allopatric Japan Sea female was collected in Lake Shinji in March 2014. The outgroup species, *G. wheatlandi* was captured from Demarest Lloyd State Park, MA, USA in 2007, as described previously [45]. Libraries were constructed with TruSeq DNA Sample Preparation Kit (Illumina) and whole-genome 100 bp paired-end sequencing was performed on an Illumina HiSeq2000 at the National Institute of Genetics (sympatric JS and PO) and Functional Genomics Facility, NIBB Core Research Facilities (allopatric JS) [45]. Whole genome sequencing of North American marine and stream populations collected from Little Campbell River, BC, Canada was reported previously [56,89]. For the six Atlantic *G. aculeatus* individuals (North Sea) included in the study, we used previously published sequences [51]. All Japan Sea, Pacific Ocean, Little Campbell and *G. wheatlandi* individuals were sequenced to a high mean depth of coverage (61X), whereas Atlantic individuals had a lower depth of 12X (see S7 Table for more information)

Japanese individuals used for RAD sequencing have been previously described elsewhere [63]. Samples used for RAD sequencing from the Atlantic lineage were collected from across Ireland in 2009–2011 [90,91]. DNA was extracted using a Qiagen DNeasy Blood and Tissue Kit (Qiagen, Valencia, CA, USA). Single digest RAD-sequencing was performed using *Sbf*I following a standard protocol [92]. RAD library preparation and sequencing was conducted using a 100bp single-end Illumina HiSeq by Floragenex (Oregon, USA).

Accession numbers, sample names and locations for all genome and RAD-seq samples are listed in S7 Table.

Whole genome alignment, variant calling and phasing. Sequence reads were mapped to the Broads S1 stickleback reference genome [93] using CLC Genomics Workbench 8.0 (Qiagen, Hilden, Germany) as described previously [45]. Alignments were exported as bam files and were sorted and indexed using samtools 1.2 [94]. We first called bases at all sites (i.e. variant and invariant) across the genome for all 27 resequenced individuals and the outgroup (*G. wheatlandi*). Mapped reads from all individuals were piled-up using samtools *mpileup* and called against the stickleback BROAD S1 reference genome using the bcftools 1.2 consensus caller, adjusting for poor mapping quality (-c 50) [95]. This consensus call produced a vcf file with a base call for every position in the genome for all samples (27 + 1 outgroup). Consensus calls from this phase were used in later demographic inference using PSMC and ABC with separate filtering criteria applied to each (see relevant sections). Estimates of f_d, d_{XY} and G_{MIN}, were also produced from this callset (see below). For analysis of differentiation and introgression we produced two separate datasets to compare the effects of phasing on our approach (see S1 Text). The first unphased dataset used only Phred quality score >20 and a maximum depth

of 200 (representing four times the mean coverage for resequenced individuals). The second dataset was phased and as such required more stringent filtering. For this we allowed a maximum of two alleles at each position (to retain invariant sites), an MAF of 0.05, a minimum call rate of 80% across all individuals, a minimum site depth of 10 and maximum site depth of 200. Phasing was conducted using ShapeIt2 using default parameters [96].

We further filtered our callset down to produce a subset of high-quality biallelic SNPs with which to examine genome-wide differentiation (measured as F_{ST}) between the Japan Sea, Pacific and Atlantic Ocean lineages. We used bcftools to filter the consensus-call vcf for these three lineages, only retaining sites with a Phred Quality score >10, and with a maximum individual read depth of 200; loci with very high coverage may represent gene duplication and are excluded. Prior to estimating F_{ST} (see below), we filtered for bi-allelic sites with an MAF > 0.05, genotype calls in at least 70% of individuals, a minimum depth of 10 and a maximum depth of 200 using vcftools.

Mitochondrial genome divergence. To estimate divergence times based on mitochondrial DNA, we performed Bayesian coalescent analysis using BEAST v2.2.1 [97]. From our resequencing data, we extracted the whole mitochondrial genome from the 26 Japan Sea, Pacific and Atlantic Ocean individuals. We also downloaded two *G. wheatlandi* whole mitogenomes as outgroups (NCBI accession numbers: AB445129 & NC011570). Note that due to poor sequence coverage across the mitogenome we excluded our own re-sequenced *G. wheatlandi* individual here. Mitogenomes were aligned using MUSCLE v3.81.3 [98] resulting in a 16,549 bp final alignment and were not partitioned for phylogenetic analysis.

Although there is a considerable three-spined stickleback fossil record, it is unfortunately of little use for providing fossil calibration dates for splits within the *Gasterosteus* genus [99,100]. However biogeographical events can also be used to calibrate node estimates and as such we used a normal prior (mean = 1.5 million years, SD = 0.75 million years) on the split between the Japan Sea and Pacific Ocean *G. aculeatus* lineages. We provided a further normal prior on the split date between the Pacific and Atlantic Ocean mitochondrial lineages (mean = 0.5 million years, SD = 0.25 million years). The latter prior distribution was intentionally made wide to reflect uncertainty surrounding this estimate. Initial analyses with BEAST indicated that marginal prior distributions for node ages did not behave as specified in the model and instead returned extremely recent divergence times with low likelihood support. This is a common bias in coalescent divergence time dating and use of a calibrated prior removed this issue [101,102]. As a result, we performed all further analyses with a calibrated Yule prior. Incorrect choice of molecular clock model can seriously bias coalescent estimates of lineage divergence times and so care must be taken to ensure the appropriate model is chosen [103,104]. We used path-sampling analysis in BEAST to estimate model marginal likelihoods for three different clock models–strict, relaxed lognormal and relaxed exponential. For each model, Markov chain Monte Carlo (MCMC) was run for 5 x 10^7 iterations, and marginal likelihoods were calculated using BEAST. We then ran the final model using two 10^8 independent MCMC runs. Runs were assessed in TRACER [105] to ensure convergence and that ESS values > 200 –i.e. the posterior was adequately sampled. Independent runs were then combined to produce posterior estimates of divergence times and substitution rates.

Nuclear phylogenetic analysis and genealogical sorting index (*gsi*). To investigate nuclear phylogenetic discordance, we constructed maximum likelihood trees from consensus sequences for non-overlapping 10, 50 and 100 kb sliding windows following Martin et al [21]. The best-fit tree was estimated for each window using RAxML with a 'GTRGAMMA' model and a random number seed [106]. Trees were classified using a custom R script available from Dryad (doi:10.5061/dryad.104g3d0) that binned trees based on whether they matched three different topologies; species, geography, ancestral–or were unresolved. For the species

category, all Atlantic, Pacific and Japan Sea individuals form separate monophyletic groups; for the geography category, Japan Sea and Pacific Ocean form a monophyletic group separate to the Atlantic Ocean; trees where the Atlantic Ocean grouped monophyletically with the Japan Sea were classed as ancestral. Trees that did not fit any of these categories were classified as unresolved. Following categorisation, trees were then standardised to ensure equal branch lengths using the *compute.brlen* function from the *Phytools* R package [107] and were finally visualised for each gene tree class using the *densiTree* function in the R package *Phangorn* [108].

We additionally used the non-overlapping Maximum Likelihood phylogenies to calculate genealogical sorting index (*gsi*) [52]. We used a custom R script to estimate *gsi* across the autosomes of 26 resequenced individuals. This allowed us to compare autosomal signals of introgression with a reduction in *gsi*.

Population size change over time. We used PSMC to estimate fluctuations in effective population size over time [109]. PSMC uses the density of heterozygote sites across a single diploid genome to estimate blocks of constant TMRCA that are split by recombination and then uses these to infer ancestral effective population sizes (N_e) over time [74,109]. Since PSMC can only analyse a single diploid individual at a time, we ran the program separately on each of the 26 resequenced genomes from Japan Sea, Pacific and Atlantic Ocean lineages. We additionally ran the analyses for a resequenced genome of a marine ecotype fish from Little Campbell River, Canada as a representative of the Eastern Pacific. Consensus sequences for each genome were converted to PSMC format—a binary format indicating the presence/absence of heterozygous sites within a specified window. We used 100 bp windows along a scaffold, requiring a minimum of 10,000 'good' sites (i.e. those passing with a Phred quality filter >20, with a minimum depth of 20 and maximum depth of 120) to be present on a genome scaffold in order for it to be included. We then ran PSMC for 30 iterations with a maximum coalescent time of 15 (measured in units of $2N_O$ where N_O is ancestral population size). Due to the difficulty of inferring past effective population sizes across this time, PSMC requires the user to provide intervals which are combined to produce the same effective population size [109]. Since this method is least accurate for recent (i.e. < 20 kyr BP) and more ancient periods [109], we estimated N_e for 45 discrete time intervals, combining the first four and the last three intervals using the command "4+19*2+3". To scale our results from coalescent units, we assumed a generation time of 1 year [110] and used an autosomal mutation rate of 7.1×10^{-9} per site per year [111]. Finally, to provide confidence intervals for our N_e estimates, we performed 100 bootstraps on 500 kb segments for each analysis.

Approximate Bayesian computation (ABC). We used ABC to test different scenarios of divergence between the Japan Sea and Pacific Ocean lineages and to estimate demographic parameters, such as divergence time and migration rate, under these scenarios.

To obtain loci suitable for our ABC analysis, we randomly sampled nuclear loci from the 20 resequenced genomes (sympatric Japan Sea and Pacific Ocean) using a similar approach to Nadachowska-Brzyska et al [17]. Using a custom R script, we produced a bed file of reference genome coordinates for 2 kb loci randomly sampled at 125 kb intervals; resulting in 2,378 potential loci per individual. Using a custom python script, we called sequences for each locus from the consensus vcf, coding heterozygous bases with IUPAC codes. This script created two haplotype sequences for each of the 2 kb loci, randomly assigning heterozygous variants to one of the two called haplotypes; this step allowed us to use unphased data for demographic analyses [66,112]. We then further filtered these loci to include only those that occur on autosomes, with >1,000 bp sequence and with a base call at each position for at least 14 of the 20 individuals (i.e. a 30% missing data threshold). This resulted in a final dataset of 1,874 loci. Functions and scripts for generating coordinates and extracting and filtering consensus sequences are

available from Dryad (doi:10.5061/dryad.104g3d0) and on Github (https://github.com/markravinet/genome_sampler).

Following Robinson et al [66] we used a custom R-based control script and msABC [113] to perform simulations, calculate summary statistics and quantify their distribution across the genome in a single step. This approach offers considerable flexibility in establishing prior probability distributions for each of the estimated parameters. Furthermore, given the large size of our dataset (i.e. approximately 2,000 loci for 20 individuals), each simulation produces a large amount of data, making storage a challenge. Using R to interface with msABC allowed us to greatly reduce the required data storage.

For each of the 15 models we performed 10^6 simulations. We used a combination of GNU Parallel [114] and independent runs across multiple computing cores to reduce analysis time to approximately 1 day per model (scripts and additional instructions available from Dryad: doi:10.5061/dryad.104g3d0) and on Github (https://github.com/markravinet/abc_pipeline).

We initially ran our simulations to produce all the available summary statistics that msABC calculates. However since summary statistic choice can greatly alter the outcomes of ABC analyses [115,116], all post-simulation ABC analyses were conducted using subsets of 29, 20 and 12 summary statistics. Following completion of the simulation step, we performed a neural-network rejection step on log-transformed parameter estimates with a tolerance of 0.01 using the abc function in the R package *abc* [117]. The neural network rejection method performs better with higher dimensionality in the data and weights the accepted summary statistics based on their distance from the observed dataset [117,118] Posterior probability was estimated for each model using the R *abc postpr* function, also with a neural-network method for a range of tolerance values representing 0.1%, 0.5%, 1% and 3% of the simulated data (i.e. 1,000, 5,000, 10,000 and 30,000 datasets respectively). With a standard rejection ABC approach, posterior probabilities of models are calculated from proportion of simulations from each model accepted after the rejection step; therefore if 1,000 simulations are accepted and all are from a single model, the posterior probability is 1 for that model and 0 for all others. Using a neural-network, the distances between the observed summary statistics and those from the simulations are weighted in a non-linear regression model, allowing a more accurate estimation of posterior probability when dimensionality in the data is high [117,119]. In keeping with a hierarchical analysis [17], we performed two rounds of model selection. We first chose the growth model with the highest posterior probability within each divergence scenario. Following this, we performed model selection on the five models with the highest support within each divergence category.

In order to ensure our ABC approach was reliable, we used pseudo-observed datasets (PODs) to assess how well we could discriminate between different divergence scenarios. Essentially, this involves randomly selecting a series of simulated dataset from a known model (hence pseudo-observed) and then rerunning the model selection procedure to see whether the true model could be recovered. For further details of our POD-based sensitivity analysis and ABC approach, see S1 Text.

Detecting genome-wide divergence and recent introgression. Weir and Cockerham's F_{ST} [119] was calculated using 10 and 50 kb non-overlapping windows with VCFtools 0.113 [120]. To calculate statistics such as d_{XY}, G_{MIN} and f_d, we used a modified version of a python script used by Martin et al [21]. In addition to our main filters on the dataset (see *Genome Alignment and Variant calling*), we only calculated these haplotype-based statistics for windows with >50% of useable bases–i.e. >5,000 sites within a 10 kb sliding window. For autosomal statistics, all individuals were included in the analyses. For comparing the ancestral (chrXIX) and neo-sex chromosomes (chrIX) with autosomes (Fig 5E and 5F), we re-ran the analyses of all chromosomes using only females. In addition to 10 kb windows, we also

performed analyses for non-overlapping 50 kb windows to aid visualisation; the results from all analyses were then combined into a single dataset using custom R scripts.

We calculated recently established statistics, G_{MIN} and f_d, for detecting introgression between divergent lineages [30,59]. G_{MIN} is particularly suited for identifying recent, low frequency introgression [30] whereas f_d can also identify stronger, high frequency introgression events [59]. Importantly, both methods are robust to variation in recombination rate variation. Initial genome scans conducted using G_{MIN} revealed a series of 'valleys' present across the genome. Detection of such valleys, like genomic islands of divergence, presents a variety of methodological issues. Firstly, how do we determine that G_{MIN} valleys are not due to stochastic variation in genealogy amongst loci? Secondly, how do we measure the size and distribution of valleys of introgression? Finally, how can we determine a null or expected distribution of valleys across the genome to test for the under- or overrepresentation of valleys?

To deal with each of these issues in turn, we first performed chromosome-specific permutations to identify the null distribution of the value of G_{MIN}. Specifically, we shuffled the nucleotide sequence of each chromosome 100 times and estimated G_{MIN} for 11 different sliding window sizes (5,000, 5,500, 6,000, 6,500, 7,000, 7,500, 8,000, 8,500, 9,000, 9,500 and 10,000 kb), representing the distribution of useable sites from the empirical dataset. We then used the lower 99 percentile of the permutations to determine the value of G_{MIN} below which a window could be classified as a valley. Identifying the boundaries of divergent genome regions is somewhat subjective and open to potential bias [121]. To account for this, we used a hidden Markov-model (HMM) approach to classify windows into two states—i.e. valleys or non-valleys–and to estimate the probability of state transition. Following Soria-Caracasco *et al.* [60], we used the R package HiddenMarkov [122] on a logit transformed G_{MIN} distribution. Transition probabilities between the two states were symmetrical with an emphasis on it being difficult to transition between states ($p = 0.1$) but relatively easy to remain within a state ($p = 0.9$). Since valleys are relatively rare in the genome, we set our models to start in the non-valley state and we provided estimated parameter values for the states based on the empirical distribution. HMM estimates were run for both the sympatric and allopatric comparisons using the *baum-welch* function to estimate parameters using the Baum-Welch algorithm and the *viterbi* function to estimate the sequence of states using the Viterbi algorithm. We used a similar approach to identify f_d peaks but we instead performed the analysis using untransformed f_d values only in the sympatric Japanese *G. aculeatus* and *G. nipponicus* comparison.

Permutation tests. In order to test for differences between allopatric and sympatric comparisons of F_{ST}, d_{XY}, f_d and G_{MIN}, we used a permutation-based independence test implemented in the R package, *coin*. To give an example of how this approach works, consider a test of whether F_{ST} is higher in allopatry versus sympatry. Estimates of F_{ST} from 10 kb windows were randomly sampled and their identity as coming from either the allopatric or sympatric case are also permutated. This creates a null distribution of Z–i.e. the expected mean difference between the two populations.

RAD-seq data processing, population structure and ancestry analysis. To complement our whole-genome resequencing data, we performed RAD sequencing on a further 151 Japan Sea, Pacific Ocean and Atlantic individuals (see S7 Table for a full breakdown). We further combined our RAD-seq dataset with previously determined RAD sequences from 93 Pacific Ocean fish sampled in North America [53], resulting in a total dataset of 244 individuals. RAD sequence reads were demultiplexed and processed using the *process_radtags* module of Stacks 1.30 [123]. All reads were trimmed to 90 bp and any read where the average Phred quality score dropped below 10 in a 9 bp sliding window was discarded. Following filtering, reads were mapped to the Roesti *et al.* [64] build of the *G. aculeatus* genome using GSNAP [124] allowing a maximum of two indels to be present in an alignment, reporting no suboptimal

hits, allowing a maximum of 8 mismatches and printing only the best alignment. SNPs were then called using the samtools and bcftools pipeline [125]. Called variants were then filtered using vcftools to remove all sites with greater than 25% missing data, to include genotypes only with an individual depth between 15X and 100X, to remove all sites with a Phred quality score below 20 and with a minor allele frequency below 0.05. Since common admixture analyses assume independence among sites (i.e. the absence of linkage disequilibrium) [126], we additionally pruned our RAD-derived SNP dataset using plink [127], removing all sites where pairwise linkage disequilibrium was greater than 0.4 within a 100 kb window.

PCA on allele frequencies from all individuals was conducted using the *glPca* function from the R package adegenet [128]. Admixture analysis was carried out on all 244 individuals using STRUCTURE [54,55]. For each value of *K* from 1 to 8, the program was run for 10 iterations with a burn-in of 10,000 steps followed by 20,000 MCMC steps. The most likely value of *K* was assessed using STRUCTURE HARVESTER [129].

To further investigate variation in individual ancestry, we used a genomic cline approach with the R package *introgress* [130]. As our resequencing data was taken from individuals previously identified of being of probable 'pure' descent inferred by microsatellite data, we identified ancestry informative markers from this resequence dataset. To be informative, a marker was required to be present in the RAD-seq data, occur on an autosome and to have an absolute allele frequency difference of >0.8 between the two parental species (following Larson et al. [131]). For each individual, we then calculated hybrid index and interspecific heterozygosity [130]. As a measure of a power of this approach, we used *adegenet* to simulate F_1 and F_2 hybrids, as well as Japan Sea and Pacific Ocean backcrosses.

As independent support of our demographic inference using ABC, we also used a maximum likelihood inference of demography based on the joint site-frequency spectrum from Japan Sea and Pacific Ocean RAD-seq data ($N = 51$). To account for missing data, we resampled 20 genotypes per species at each site, resulting in calls for 20 'pseudo-individuals' at 22,065 SNP loci. We used the same models as the ABC analysis (without population growth parameters) but with parameters drawn from a loguniform distribution (see S1 Text for more details on parameters, models and data used). We performed 100 independent runs of 100,000 coalescent simulations for each model using fastsimcoal2 [132]. Model selection was carried out on the run with the highest likelihood using Akaike's Information Criterion (AIC); however, as our SFS dataset was not pruned for linked sites (i.e. SNPs are not independent), AIC values should be interpreted carefully [133]. As an additional mean of model selection, we also calculated the likelihood distribution for each model using 100 expected site frequency spectra and 10^6 coalescent simulations [24].

Detecting the direction of introgression. We investigated the direction of gene flow between the Japan Sea and Pacific Ocean lineages using partitioned *D* statistics [61]. This is conceptually similar to standard four population ABBA-BABA tests for gene flow but includes a fifth population–an allopatric lineage of the Japan Sea. This balances the assumed phylogeny (i.e. ((P1, P2), ($P3_1$, $P3_2$), O) and therefore allows us to rotate the populations used in the analysis–i.e. testing for an enrichment of gene flow in both directions. We therefore tested two topologies ((AT, PO), (JS_S, JS_A), O) and ((JS_A, JS_S), (PO, AT), O) (see S15 Fig). For either test topology, an excess of the ABBAA (compared to BABAA) or ABBBA (compared to BABBA) in a genome window inflates partitioned *D* statistics above zero–indicating gene flow from the P3 into P2.

Given that the partitioned *D* approach has attracted some criticisms, we also calculated D_{FOIL} statistics [62]. D_{FOIL} is an additional extension of the four population test but one that incorporates all possible introgression events for a symmetric four population tree (excluding the outgroup) (see S16 Fig). We used the same test phylogeny as with the partitioned *D* statistics.

Both partitioned D and D_{FOIL} are based on ABBA/BABA methods–i.e. where only a single individual is present at the tips of the phylogeny. To account for this, we extended both methods to account for allele frequency data, meaning our site pattern counts are weighted by allele frequencies [59]. To calculate both D and D_{FOIL} statistics, we used a modified version of a python script used by Martin et al [21].

Characterization of introgression sites. In order to characterize regions of introgression, we identified candidate regions showing a strong signature of introgression (i.e. G_{MIN} valleys and f_d peaks) from our genome scan approach. We then counted the number of unique genes falling within our candidate valleys/peaks and compared this to a null distribution generated by 1,000 random samples of 10 kb non-valley/non-peak genome windows for the same number and size range as the valleys or peaks.

We next tested whether genes in introgressed regions were more likely to have any specific functions. To achieve this, we used gene ontology (GO) analysis on genes in valleys and 1,000 randomly chosen from across the genome. GO analysis was performed with the ClueGO plugin [133] for Cytoscape 3.4.0 [134]. Since functional annotations for this analysis were drawn from the human genome, we first generated a list of human-stickleback orthologous gene IDs (Ensembl Biomart 86). We then subset our candidate and random gene sets to include only orthologous genes. Several human genes have multiple stickleback orthologs; we therefore allowed only a single, randomly chosen occurrence of each human gene in both sets to prevent pseudo-replication. A hypergeometric test was conducted for testing enrichment with Benjamini & Hochberg FDR correction [135].

Supporting information

S1 Table. Classification of trees. Proportion of trees drawn from 10, 50 and 100 kb windows representing species, ancestral and geographical topologies.
(XLSX)

S2 Table. Mitogenome analysis. Marginal log-likelihood values and Bayes factor comparisons from path sampling of substitution clock models used for mitogenome phylogeny and divergence time estimation.
(XLSX)

S3 Table. Median, L95% and U95% HPD (highest probability densities) for demographic parameters estimated under IM + bottleneck model.
(CSV)

S4 Table. Number of valleys and valley per Mb for all chromosomes.
(TXT)

S5 Table. Enriched GO terms for genes present in G_{MIN} valleys.
(XLS)

S6 Table. Enriched GO terms for genes present in f_d peaks.
(XLS)

S7 Table. Sample names, locations and sequence data accession numbers.
(XLSX)

S1 Fig. Global species distribution. The global distribution of the Pacific and Atlantic Ocean lineages of *G. aculeatus* allow sympatric and allopatric comparisons with *G. nipponicus*; AT = Atlantic Ocean, PO = Pacific Ocean and JS = Japan Sea.
(PDF)

S2 Fig. Japanese stickleback mitonuclear discordance. (A) Mitogenome Bayesian tree shows divergence between two mitochondrial clades–the Transpacific and European North American; asterisks on nodes indicate appropriate densities shown in (B). (B) Posterior probability densities for mitochondrial divergence time between *G. aculeatus* and *G. nipponicus* (pink) and between Pacific and Atlantic populations of the European North American clade (blue).
(PDF)

S3 Fig. Bootstrapped PSMC curves for 26 resequenced individuals.
(PDF)

S4 Fig. PSMC profile for all 26 individuals and an additional Eastern Pacific individual from Little Campbell River, Canada.
(PDF)

S5 Fig. Principal component analysis on RAD-seq data from 295 individuals from across the distribution of all three lineages. The arrow indicates the presence of an admixed individual occurring in the Akkeshi system.
(PDF)

S6 Fig. Structure analysis on RAD-seq data from multiple Japan Sea, Atlantic and Pacific Ocean populations. Analysis with $K = 2$ & 4 clusters (A), which is supported by likelihood analysis (B), showed the presence of admixed individuals in the Akkeshi system.
(PDF)

S7 Fig. Individual ancestry estimates using hybrid index and interspecific ancestry based on RAD-seq data.
(PDF)

S8 Fig. Genome-wide F_{ST} measured in non-overlapping 50 kb windows for allopatric and sympatric between and within species comparisons.
(PDF)

S9 Fig. Genome-wide d_{XY} measured in non-overlapping 50 kb windows for allopatric and sympatric between and within species comparisons.
(PDF)

S10 Fig. Negative association between f_d and G_{MIN} in sympatric (JS vs PO) and allopatric (JS v AT) between species comparisons.
(PDF)

S11 Fig. Genome-wide G_{MIN} for sympatric between species comparisons. Black line represents 50 kb non-overlapping window G_{MIN} signature. Points represent 10 kb windows; grey points are non-valley windows, blue points are valley windows identified by Hidden Markov Model algorithm.
(PDF)

S12 Fig. Absolute divergence (d_{XY}) is lower and f_d is higher in G_{MIN} valleys compared to the genome-wide background.
(PDF)

S13 Fig. Genome-wide f_d for sympatric between species comparisons. Black line represents 50 kb non-overlapping window f_d signature. Points represent 10 kb windows; grey points are non-valley windows, while blue points are peak windows identified by Hidden Markov Model

algorithm.
(PDF)

S14 Fig. Comparison of d_{XY} between f_d peaks and non-peaks.
(PDF)

S15 Fig. Analysis of partitioned *D* statistics. (A) Boxplots comparing partitioned *D* statistics between f_d peaks and the autosomal background. Dashed line at zero indicates a balance between allele patterns indicative of incomplete lineage sorting. In (B), P1 = Atlantic Ocean (AT), P2 = Pacific Ocean (PO), $P3_1$ = sympatric Japan Sea (JS_S), $P3_2$ = allopatric Japan Sea (JS_A), and O = *G. wheatlandi* (WT). In (C), P12 and P3 were swapped. D_1 measures asymmetry between P1 and P2 where the derived allele B is present in $P3_1$ but not $P3_2$, D_2 measures where allele B is present in $P3_2$ but not $P3_1$, and D_{12} measures where the derived allele is shared by both $P3_1$ and $P3_2$. If we assume that the derived allele B occurred at the ancestor of P3, D12 indicates introgression from P3 to P2. See [61] for a more detailed explanation of these statistics.
(PDF)

S16 Fig. D_{FOIL} statistics. Using these statistics, we assume that the divergence time between P_1 and P_2 is younger than that between P_3 and P_4 and infer all possible introgressions including ancestral introgression involving P_{12}. The number of loci (100 kb-window) that show statistically significant introgression are shown. See [62] for a more detailed explanation of these statistics.
(PDF)

S17 Fig. The relationship between introgression measured as G_{MIN} and f_d and $\log_{10}$ recombination rate.
(PDF)

S18 Fig. Venn diagram showing overlap between genes occurring in introgressed regions identified using different measures.
(PDF)

S1 Text. ABC analysis, SFS estimation and data filtering sensitivity.
(PDF)

Acknowledgments

We are grateful to Manabu Kume, Seiichi Mori, and staff at the Aquarium Gobius for providing samples and Katsushi Yamaguchi for technical assistance. Keisuke Honda, the Institute of Statistical Mathematics and the DNA Data Bank of Japan are thanked for their help with running analyses on supercomputers. We are additionally grateful to Simon Martin, John Robinson, Joana Meier and David Marques for sharing scripts and their advice on analyses. Freddy Chain and Philine Feulner are also thanked for their assistance with their sequence data. We thank Cassandra Trier for her assistance and advice with GO analyses. All members of the Kitano Lab provided invaluable advice throughout the project. We would also like to thank Mark Kirkpatrick and members of his lab and three anonymous reviewers for comments on an earlier version of this manuscript.

Author Contributions

Conceptualization: Mark Ravinet, Jun Kitano.

Data curation: Mark Ravinet, Kohta Yoshida, Shuji Shigenobu, Atsushi Toyoda, Asao Fujiyama, Jun Kitano.

Formal analysis: Mark Ravinet.

Funding acquisition: Jun Kitano.

Investigation: Mark Ravinet.

Methodology: Mark Ravinet.

Project administration: Jun Kitano.

Supervision: Jun Kitano.

Writing – original draft: Mark Ravinet, Jun Kitano.

Writing – review & editing: Kohta Yoshida.

References

1. Mayr E. Animal species and evolution. Cambridge, Massachusetts: Havard University Press; 1963.
2. Coyne JA, Orr HA. Speciation. New York: Sinaeur; 2004.
3. Nosil P. Ecological Speciation. Oxford, UK: Oxford University Press; 2012.
4. Wu C-I. The genic view of the process of speciation. J Evol Biol. 2001; 14: 851–865.
5. Feder JL, Egan SP, Nosil P. The genomics of speciation-with-gene-flow. Trends Genet. Elsevier Ltd; 2012; 28: 342–350. https://doi.org/10.1016/j.tig.2012.03.009 PMID: 22520730
6. Nosil P, Feder JL. Genomic divergence during speciation: causes and consequences. Philos Trans R Soc London Ser B. 2012; 367: 332–342. https://doi.org/10.1098/rstb.2011.0263 PMID: 22201163
7. Ravinet M, Faria R, Butlin RK, Galindo J, Bierne N, Rafajlović M, et al. Interpreting the genomic landscape of speciation: finding barriers to gene flow. J Evol Biol. 2017; 30: 1450–1477.
8. Nosil P, Feder JL, Flaxman SM, Gompert Z. Tipping points in the dynamics of speciation. Nat Ecol Evol. Macmillan Publishers Limited; 2017; 1: 1–8. https://doi.org/10.1038/s41559-016-0001 PMID: 28812620
9. Riesch R, Muschick M, Lindtke D, Villoutreix R, Comeault AA, Farkas TE, et al. Transitions between phases of genomic differentiation during stick-insect speciation. Nat Ecol Evol. Macmillan Publishers Limited, part of Springer Nature.; 2017; 1: 82. https://doi.org/10.1038/s41559-017-0082 PMID: 28812654
10. Feder JL, Nosil P, Wacholder AC, Egan SP, Berlocher SH, Flaxman SM. Genome-wide congealing and rapid transitions across the speciation continuum during speciation with gene flow. J Hered. 2014; 105: 810–820. https://doi.org/10.1093/jhered/esu038 PMID: 25149256
11. Seehausen O, Butlin RK, Keller I, Wagner CE, Boughman JW, Hohenlohe P a, et al. Genomics and the origin of species. Nat Rev Genet. Nature Publishing Group; 2014; 15: 176–92. https://doi.org/10.1038/nrg3644 PMID: 24535286
12. Wolf JB, Ellegren H. Making sense of genomic islands of differentiation in light of speciation. Nat Rev Genet. Nature Publishing Group; 2017; 18: 87–100. https://doi.org/10.1038/nrg.2016.133 PMID: 27840429
13. Carneiro M, Albert FW, Afonso S, Pereira RJ, Burbano H, Campos R, et al. The genomic architecture of population divergence between subspecies of the European Rabbit. PLoS Genet. 2014;10. https://doi.org/10.1371/journal.pgen.1003519 PMID: 25166595
14. McGaugh SE, Noor MAF. Genomic impacts of chromosomal inversions in parapatric *Drosophila* species. Philos Trans R Soc Lond B Biol Sci. 2012; 367: 422–9. https://doi.org/10.1098/rstb.2011.0250 PMID: 22201171
15. Renaut S, Grassa CJ, Yeaman S, Moyers BT, Lai Z, Kane NC, et al. Genomic islands of divergence are not affected by geography of speciation in sunflowers. Nat Commun. Nature Publishing Group; 2013; 4: 1827. https://doi.org/10.1038/ncomms2833 PMID: 23652015
16. Gagnaire PA, Pavey SA, Normandeau E, Bernatchez L. The genetic architecture of reproductive isolation during speciation-with-gene-flow in lake whitefish species pairs assessed by RAD sequencing. Evolution (N Y). 2013; 67: 2483–2497. https://doi.org/10.1111/evo.12075 PMID: 24033162
17. Nadachowska-Brzyska K, Burri R, Olason PI, Kawakami T, Smeds L, Ellegren H. Demographic Divergence History of Pied Flycatcher and Collared Flycatcher Inferred from Whole-Genome Re-

sequencing Data. Payseur BA, editor. PLoS Genet. 2013; 9: e1003942. https://doi.org/10.1371/journal.pgen.1003942 PMID: 24244198

18. Burri R, Nater A, Kawakami T, Mugal CF, Olason PI, Smeds L, et al. Linked selection and recombination rate variation drive the evolution of the genomic landscape of differentiation across the speciation continuum of Ficedula flycatchers. Genome Res. 2015; 25: 1656–1665. https://doi.org/10.1101/gr.196485.115 PMID: 26355005

19. Phifer-Rixey M, Bomhoff M, Nachman MW. Genome-wide patterns of differentiation among house mouse subspecies. Genetics. 2014; 198: 283–297. https://doi.org/10.1534/genetics.114.166827 PMID: 24996909

20. Kenney AM, Sweigart AL. Reproductive isolation and introgression between sympatric Mimulus species. Mol Ecol. 2016; 25: 2499–2517. https://doi.org/10.1111/mec.13630 PMID: 27038381

21. Martin SH, Dasmahapatra KK, Nadeau NJ, Salazar C, Walters JR, Simpson F, et al. Genome-wide evidence for speciation with gene flow in Heliconius butterflies. Genome Res. 2013; 23: 1817–28. https://doi.org/10.1101/gr.159426.113 PMID: 24045163

22. Martin SH, Eriksson A, Kozak KM, Manica A, Jiggins CD. Speciation in Heliconius Butterflies: Minimal Contact Followed by Millions of Generations of Hybridisation. BioRxiv. 2015; 1–24. https://doi.org/10.1101/015800

23. Bierne N, Gagnaire PA, David P. The geography of introgression in a patchy environment and the thorn in the side of ecological speciation. Curr Zool. 2013; 59: 72–86.

24. Meier JI, Sousa VC, Marques DA, Selz OM, Wagner CE, Excoffier L, et al. Demographic modelling with whole-genome data reveals parallel origin of similar Pundamilia cichlid species after hybridization. Mol Ecol. 2017; 26: 123–141. https://doi.org/10.1111/mec.13838 PMID: 27613570

25. Cruickshank TE, Hahn MW. Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. Mol Ecol. 2014; 23: 3133–3157. https://doi.org/10.1111/mec.12796 PMID: 24845075

26. Noor MAF, Bennett SM. Islands of speciation or mirages in the desert? Examining the role of restricted recombination in maintaining species. Heredity (Edinb). 2009; 103: 439–444.

27. Rosenzweig BK, Pease JB, Besansky NJ, Hahn MW. Powerful methods for detecting introgressed regions from population genomic data. Mol Ecol. 2016; 25: 2387–2397. https://doi.org/10.1111/mec.13610 PMID: 26945783

28. Nater A, Burri R, Kawakami T, Smeds L, Ellegren H. Resolving evolutionary relationships in closely related species with whole-genome sequencing data. Syst Biol. 2015; 64: 1000–1017. https://doi.org/10.1093/sysbio/syv045 PMID: 26187295

29. Geneva A, Garrigan D. Population Genomics of Secondary Contact. Genes (Basel). 2010; 1: 124–142. https://doi.org/10.3390/genes1010124 PMID: 24710014

30. Geneva AJ, Muirhead CA, Kingan SB, Garrigan D. A new method to scan genomes for introgression in a secondary contact model. PLoS One. 2015; 10: e0118621. https://doi.org/10.1371/journal.pone.0118621 PMID: 25874895

31. Meier JI, Marques DA, Mwaiko S, Wagner CE, Excoffier L, Seehausen O. Ancient hybridization fuels rapid cichlid fish adaptive radiations. Nat Commun. Nature Publishing Group; 2017; 8: 14363. https://doi.org/10.1038/ncomms14363 PMID: 28186104

32. Martin CH, Crawford JE, Turner BJ, Simons LH. Diabolical survival in Death Valley: recent pupfish colonization, gene flow and genetic assimilation in the smallest species range on earth. Proc R Soc B Biol Sci. 2016; 283: 20152334. https://doi.org/10.1098/rspb.2015.2334 PMID: 26817777

33. Hendry AP, Bolnick DI, Berner D, Peichel CL. Along the speciation continuum in sticklebacks. J Fish Biol. 2009; 75: 2000–2036. https://doi.org/10.1111/j.1095-8649.2009.02419.x PMID: 20738669

34. McKinnon JS, Rundle HD. Speciation in nature: the threespine stickleback model systems. Trends Ecol Evol. 2002; 17: 480–481. https://doi.org/10.1016/S0169-5347(02)02579-X

35. Feulner PGD, Chain FJJ, Panchal M, Huang Y, Eizaguirre C, Kalbe M, et al. Genomics of Divergence along a Continuum of Parapatric Population Differentiation. PLOS Genet. 2015; 11: e1004966. https://doi.org/10.1371/journal.pgen.1004966 PMID: 25679225

36. Roesti M, Hendry AP, Salzburger W, Berner D. Genome divergence during evolutionary diversification as revealed in replicate lake-stream stickleback population pairs. Mol Ecol. 2012; 21: 2852–2862. https://doi.org/10.1111/j.1365-294X.2012.05509.x PMID: 22384978

37. Roesti M, Kueng B, Moser D, Berner D. The genomics of ecological vicariance in threespine stickleback fish. Nat Commun. Nature Publishing Group; 2015; 6: 8767. https://doi.org/10.1038/ncomms9767 PMID: 26556609

38. Marques DA, Lucek K, Meier JI, Mwaiko S, Wagner CE, Excoffier L, et al. Genomics of Rapid Incipient Speciation in Sympatric Threespine Stickleback. PLOS Genet. 2016; 12: e1005887. https://doi.org/10.1371/journal.pgen.1005887 PMID: 26925837

39. Wu C-I, Ting C-T. Genes and speciation. Nat Rev Genet. 2004; 5: 114–22. https://doi.org/10.1038/nrg1269 PMID: 14735122

40. Nosil P, Harmon LJ, Seehausen O. Ecological explanations for (incomplete) speciation. Trends Ecol Evol. 2009; 24: 145–156. https://doi.org/10.1016/j.tree.2008.10.011 PMID: 19185951

41. Kitano J, Ross JA, Mori S, Kume M, Jones FC, Chan YF, et al. A role for neo-sex chromosomes in stickleback speciation. Nature. 2009; 461: 1079–1083. https://doi.org/10.1038/nature08441 PMID: 19783981

42. Kitano J, Mori S, Peichel CL. Phenotypic divergence and reproductive isolation between sympatric forms of Japanese threespine sticklebacks. Biol J Linn Soc. 2007; 91: 671–685. https://doi.org/10.1111/j.1095-8312.2007.00824.x

43. Higuchi M, Sakai H, Goto A. A new threespine stickleback, Gasterosteus nipponicus sp. nov. (Teleostei: Gasterosteidae), from the Japan Sea region. Ichthyol Res. 2014; 1–2. https://doi.org/10.1007/s10228-014-0403-1

44. Higuchi M, Goto A. Genetic evidence supporting the existence of two distinct species in the genus *Gasterosteus* around Japan. Environ Biol Fishes. 1996; 47: 1–16. https://doi.org/10.1007/BF00002375

45. Yoshida K, Makino T, Yamaguchi K, Shigenobu S, Hasebe M, Kawata M, et al. Sex Chromosome Turnover Contributes to Genomic Divergence between Incipient Stickleback Species. Zhang J, editor. PLoS Genet. 2014; 10: e1004223. https://doi.org/10.1371/journal.pgen.1004223 PMID: 24625862

46. Kume M, Kitano J, Mori S, Shibuya T. Ecological divergence and habitat isolation between two migratory forms of Japanese threespine stickleback (*Gasterosteus aculeatus*). J Evol Biol. 2010; 23: 1436–1446. https://doi.org/10.1111/j.1420-9101.2010.02009.x PMID: 20456572

47. Ravinet M, Takeuchi N, Kume M, Mori S, Kitano J. Comparative Analysis of Japanese Three-Spined Stickleback Clades Reveals the Pacific Ocean Lineage Has Adapted to Freshwater Environments while the Japan Sea Has Not. Craft JA, editor. PLoS One. 2014; 9: e112404. https://doi.org/10.1371/journal.pone.0112404 PMID: 25460163

48. Lackey AC, Boughman JW. Evolution of reproductive isolation in stickleback fish. Evolution (N Y). 2017; 71: 357–371.

49. Ortí G, Bell MA, Reimchen TE, Meyer A. Global survey of mitochondrial DNA sequences in the threespine stickleback: evidence for recent migrations. Evolution (N Y). 1994; 48: 608–622.

50. Yamada M, Higuchi M, Goto A. Extensive introgression of mitochondrial DNA found between two genetically divergent forms of threespine stickleback, *Gasterosteus aculeatus*, around Japan. Environ Biol Fishes. 2001; 61: 269–284.

51. Feulner PGD, Chain FJJ, Panchal M, Eizaguirre C, Kalbe M, Lenz TL, et al. Genome-wide patterns of standing genetic variation in a marine population of three-spined sticklebacks. Mol Ecol. 2012; no-no. https://doi.org/10.1111/j.1365-294X.2012.05680.x PMID: 22747593

52. Cummings MP, Neel MC, Shaw KL. A genealogical approach to quantifying lineage divergence. Evolution. 2008; 62: 2411–22. https://doi.org/10.1111/j.1558-5646.2008.00442.x PMID: 18564377

53. Catchen J, Bassham S, Wilson T, Currey M, O'Brien C, Yeates Q, et al. The population structure and recent colonization history of Oregon threespine stickleback determined using restriction-site associated DNA-sequencing. Mol Ecol. 2013; 22: 2864–2883. https://doi.org/10.1111/mec.12330 PMID: 23718143

54. Falush D, Stephens M, Pritchard JK. Inference of population structure: Extensions to linked loci and correlated allele frequencies. Genetics. 2003; 164: 1567–1587. PMID: 12930761

55. Pritchard JK, Stephens M, Rosenberg NA, Donnelly P. Association mapping in structured populations. Am J Hum Genet. 2000;67.

56. Kusakabe M, Ishikawa A, Ravinet M, Yoshida K, Makino T, Toyoda A, et al. Genetic basis for variation in salinity tolerance between stickleback ecotypes. Mol Ecol. 2016; https://doi.org/10.1111/mec.13875 PMID: 27706866

57. Hohenlohe PA, Bassham S, Etter PD, Stiffler N, Johnson EA, Cresko WA. Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. PLoS Genet. 2010; 6: e1000862. https://doi.org/10.1371/journal.pgen.1000862 PMID: 20195501

58. Samuk K, Owens GL, Delmore KE, Miller S, Rennison DJ, Schluter D. Gene flow and selection interact to promote adaptive divergence in regions of low recombination. Mol Ecol. 2017; 1–13.

59. Martin SH, Davey JW, Jiggins CD. Evaluating the use of ABBA-BABA statistics to locate introgressed loci. Mol Biol Evol. 2014; 32: 244–257. https://doi.org/10.1093/molbev/msu269 PMID: 25246699

60. Soria-Carrasco V, Gompert Z, Comeault AA, Farkas TE, Parchman TL, Johnston JS, et al. Stick Insect Genomes Reveal Natural Selection's Role in Parallel Speciation. Science. 2014; 344: 738–742. https://doi.org/10.1126/science.1252136 PMID: 24833390

61. Eaton DAR, Ree RH. Inferring phylogeny and introgression using RADseq data: An example from flowering plants (Pedicularis: Orobanchaceae). Syst Biol. 2013; 62: 689–706. https://doi.org/10.1093/sysbio/syt032 PMID: 23652346

62. Pease JB, Hahn MW. Detection and Polarization of Introgression in a Five-Taxon Phylogeny. Syst Biol. 2015; 64: 651–662. https://doi.org/10.1093/sysbio/syv023 PMID: 25888025

63. Cassidy L, Ravinet M, Mori S, Kitano J. Are Japanese freshwater populations of threespine stickleback derived from the Pacific Ocean lineage? Evol Ecol Res. 2013; 15: 295–311.

64. Roesti M, Moser D, Berner D. Recombination in the threespine stickleback genome—Patterns and consequences. Mol Ecol. 2013; 22: 3014–3027. https://doi.org/10.1111/mec.12322 PMID: 23601112

65. Knowles LL. Statistical Phylogeography. Annu Rev Ecol Syst. 2009; 40: 593–612.

66. Robinson JD, Bunnefeld L, Hearn J, Stone GN, Hickerson MJ. ABC inference of multi-population divergence with admixture from un-phased population genomic data. Mol Ecol. 2014; 23: 4458–4471. https://doi.org/10.1111/mec.12881 PMID: 25113024

67. Sousa V, Hey J. Understanding the origin of species with genome-scale data: modelling gene flow. Nat Rev Genet. Nature Publishing Group; 2013; 14: 404–14. https://doi.org/10.1038/nrg3446 PMID: 23657479

68. Mandeville EG, Parchman TL, Thompson KG, Compton RI, Gelwicks KR, Song SJ, et al. Inconsistent reproductive isolation revealed by interactions between *Catostomus* fish species. Evol Lett. 2017; 255–268. https://doi.org/10.1002/evl3.29

69. Scascitelli M, Whitney KD, Randell R a., King M, Buerkle C a., Rieseberg LH. Genome scan of hybridizing sunflowers from Texas (Helianthus annus and H. debilis) reveals asymmetric patterns of introgression and small islands of genomic differentiation. Mol Ecol. 2010; 19: 521–541. https://doi.org/10.1111/j.1365-294X.2009.04504.x PMID: 20355258

70. Lexer C, Joseph J a, van Loo M, Barbará T, Heinze B, Bartha D, et al. Genomic admixture analysis in European Populus spp. reveals unexpected patterns of reproductive isolation and mating. Genetics. 2010; 186: 699–712. https://doi.org/10.1534/genetics.110.118828 PMID: 20679517

71. Yamada M, Higuchi M, Goto A. Long-term occurrence of hybrids between Japan Sea and Pacific Ocean forms of threespine stickleback, *Gasterosteus aculeatus*, in Hokkaido Island, Japan. Environ Biol Fishes. 2007; 80: 435–443.

72. Toews DPL, Brelsford A. The biogeography of mitochondrial and nuclear discordance in animals. Mol Ecol. 2012; 21: 3907–30. https://doi.org/10.1111/j.1365-294X.2012.05664.x PMID: 22738314

73. Li H, Durbin R. Inference of human population history from individual whole-genome sequences. Nature. Nature Publishing Group; 2011; 475: 493–6. https://doi.org/10.1038/nature10231 PMID: 21753753

74. Schiffels S, Durbin R. Inferring human population size and separation history from multiple genome sequences. Nat Genet. Nature Publishing Group; 2014; 46: 919–925. https://doi.org/10.1038/ng.3015 PMID: 24952747

75. Nadachowska-Brzyska K, Burri R, Smeds L, Ellegren H. PSMC-analysis of effective population sizes in molecular ecology and its application to black-and-white *Ficedula* flycatchers. Mol Ecol. 2016; n/a-n/a. https://doi.org/10.1111/mec.13540 PMID: 26797914

76. Wirtz P. Mother species–father species: unidirectional hybridization in animals with female choice. Anim Behav. 1999; 58: 1–12. https://doi.org/10.1006/anbe.1999.1144 PMID: 10413535

77. Lescak EA, Bassham SL, Catchen J, Gelmond O, Sherbick ML, von Hippel FA, et al. Evolution of stickleback in 50 years on earthquake-uplifted islands. Proc Natl Acad Sci U S A. 2015; 112: E7204–E7212. https://doi.org/10.1073/pnas.1512020112 PMID: 26668399

78. Kume M, Kitamura T, Takahashi H, Goto A. Distinct spawning migration patterns in sympatric Japan Sea and Pacific Ocean forms of threespine stickleback *Gasterosteus aculeatus*. Icthyological Res. 2005; 52: 189–193. https://doi.org/10.1007/s10228-005-0269-3

79. Flaxman SM, Wacholder AC, Feder JL, Nosil P. Theoretical models of the influence of genomic architecture on the dynamics of speciation. Mol Ecol. 2014; 23: 4074–4088. https://doi.org/10.1111/mec.12750 PMID: 24724861

80. Barton NH, de Cara MAR. The evolution of strong reproductive isolation. Evolution. 2009; 63: 1171–90. https://doi.org/10.1111/j.1558-5646.2009.00622.x PMID: 19154394

81. Barton N, Bengtsson BO. The barrier to genetic exchange between hybridising populations. Heredity (Edinb). 1986; 57: 357–376. https://doi.org/10.1038/hdy.1986.135

82. Muirhead CA, Presgraves DC. Hybrid Incompatibilities, Local Adaptation, and the Genomic Distribution of Natural Introgression between Species. Am Nat. 2016; 187: 249–261. https://doi.org/10.1086/684583 PMID: 26807751

83. Hedrick PW. Adaptive introgression in animals: examples and comparison to new mutation and standing variation as sources of adaptive variation. Mol Ecol. 2013; https://doi.org/10.1111/mec.12415 PMID: 23906376

84. Racimo F, Sankararaman S, Nielsen R, Huerta-Sánchez E. Evidence for archaic adaptive introgression in humans. Nat Rev Genet. Nature Publishing Group; 2015; 16: 359–371. https://doi.org/10.1038/nrg3936 PMID: 25963373

85. Castric V, Bechsgaard J, Schierup MH, Vekemans X. Repeated adaptive introgression at a gene under multiallelic balancing selection. PLoS Genet. 2008; 4. https://doi.org/10.1371/journal.pgen.1000168 PMID: 18769722

86. Elgvin TO, Trier CN, Tørresen OK, Hagen, Ingerid J, Lien S, Nederbragt AJ, et al. The genomic mosaicism of hybrid speciation. Sci Adv. 2017; 3.

87. Sun C, Huo D, Southard C, Nemesure B, Hennis A, Cristina Leske M, et al. A signature of balancing selection in the region upstream to the human UGT2B4 gene and implications for breast cancer risk. Hum Genet. 2011; 130: 767–775. https://doi.org/10.1007/s00439-011-1025-6 PMID: 21660508

88. Bolnick DI, Stutz WE. Frequency dependence limits divergent evolution by favouring rare immigrants over residents. Nature. Nature Publishing Group; 2017; 546: 285–288. https://doi.org/10.1038/nature22351 PMID: 28562593

89. Ishikawa A, Kusakabe M, Yoshida K, Ravinet M, Makino T, Toyoda A, et al. Different contributions of local- and distant-regulatory changes to transcriptome divergence between stickleback ecotypes. Evolution. 2017; 71: 565–581. https://doi.org/10.1111/evo.13175 PMID: 28075479

90. Ravinet M, Harrod C, Eizaguirre C, Prodöhl P a. Unique mitochondrial DNA lineages in Irish stickleback populations: cryptic refugium or rapid recolonization? Ecol Evol. 2013; 4: 2488–2504. https://doi.org/10.1002/ece3.853 PMID: 25360281

91. Ravinet M, Prodöhl PA, Harrod C. On Irish stickleback: morphological diversification in a secondary contact zone. Evol Ecol Res. 2013; 15: 271–294.

92. Baird N a, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis Z a, et al. Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS One. 2008; 3: e3376. https://doi.org/10.1371/journal.pone.0003376 PMID: 18852878

93. Jones FC, Grabherr MG, Chan YF, Russell P, Mauceli E, Johnson J, et al. The genomic basis of adaptive evolution in threespine sticklebacks. Nature. 2012; 484: 55–61. https://doi.org/10.1038/nature10944 PMID: 22481358

94. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009; 25: 2078–9. https://doi.org/10.1093/bioinformatics/btp352 PMID: 19505943

95. Danecek P, McCarthy SA. BCFtools/csq: Haplotype-aware variant consequences. Bioinformatics. 2017; 33: 2037–2039. https://doi.org/10.1093/bioinformatics/btx100 PMID: 28205675

96. O'Connell J, Gurdasani D, Delaneau O, Pirastu N, Ulivi S, Cocca M, et al. A General Approach for Haplotype Phasing across the Full Spectrum of Relatedness. PLoS Genet. 2014;10. https://doi.org/10.1371/journal.pgen.1004234 PMID: 24743097

97. Drummond AJ, Rambaut A. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol Biol. 2007; 7: 214–222. https://doi.org/10.1186/1471-2148-7-214 PMID: 17996036

98. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004; 32: 1792–7. https://doi.org/10.1093/nar/gkh340 PMID: 15034147

99. Bell MA. Paleobiology and evolution of threespine stickleback. In: Bell MA, Foster SA, editors. The evolutionary biology of the threespine stickleback. Oxford: Oxford University Press; 1994. pp. 438–471.

100. Bell MA, Stewart JD, Park PJ. The world's oldest fossil threespine stickleback fish. Copeia. 2009; 2009: 256–265.

101. Heled J, Drummond AJ. Calibrated tree priors for relaxed phylogenetics and divergence time estimation. Syst Biol. 2012; 61: 138–49. https://doi.org/10.1093/sysbio/syr087 PMID: 21856631

102. Wheat CW, Wahlberg N. Critiquing blind dating: the dangers of over-confident date estimates in comparative genomics. Trends Ecol Evol. Elsevier Ltd; 2013; 28: 636–42. https://doi.org/10.1016/j.tree.2013.07.007 PMID: 23973265

103. Baele G, Lemey P, Bedford T, Rambaut A, Suchard M a, Alekseyenko A V. Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty. Mol Biol Evol. 2012; 29: 2157–67. https://doi.org/10.1093/molbev/mss084 PMID: 22403239

104. Baele G, Li WLS, Drummond AJ, Suchard, Lemey P. Accurate model selection of relaxed molecular clocks in bayesian phylogenetics. Mol Biol Evol. 2013; 30: 239–43. https://doi.org/10.1093/molbev/mss243 PMID: 23090976

105. Rambaut A, Drummond AJ. Tracer v1.5. 2009.

106. Stamatakis A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics. 2006; 22: 2688–2690. https://doi.org/10.1093/bioinformatics/btl446 PMID: 16928733

107. Revell LJ. phytools: an R package for phylogenetic comparative biology (and other things). Methods Ecol Evol. 2012; 3: 217–223. https://doi.org/10.1111/j.2041-210X.2011.00169.x

108. Schliep KP. phangorn: Phylogenetic analysis in R. Bioinformatics. 2011; 27: 592–593. https://doi.org/10.1093/bioinformatics/btq706 PMID: 21169378

109. Li H, Durbin R. Inference of human population history from individual whole-genome sequences. Nature. 2011; 475: 493–6. https://doi.org/10.1038/nature10231 PMID: 21753753

110. Bell MA, Foster SA. Introduction to the evolutionary biology of the threespine stickleback. In: Bell MA, Foster SA, editors. The Evolutionary Biology of the Threespine Stickleback. Oxford: Oxford University Press; 1994. pp. 1–27.

111. Guo B, Chain FJ, Bornberg-Bauer E, Leder EH, Merilä J. Genomic divergence between nine- and three-spined sticklebacks. BMC Genomics. 2013; 14: 756. https://doi.org/10.1186/1471-2164-14-756 PMID: 24188282

112. Mailund T, Halager AE, Westergaard M, Dutheil JY, Munch K, Andersen LN, et al. A new isolation with migration model along complete genomes infers very different divergence processes among closely related great ape species. PLoS Genet. 2012; 8: e1003125. https://doi.org/10.1371/journal.pgen.1003125 PMID: 23284294

113. Pavlidis P, Laurent S, Stephan W. msABC: a modification of Hudson's ms to facilitate multi-locus ABC analysis. Mol Ecol Resour. 2010; 10: 723–7. https://doi.org/10.1111/j.1755-0998.2010.02832.x PMID: 21565078

114. Tange O. GNU Parallel—The Command-Line Power Tool.;login USENIX Mag. Frederiksberg, Denmark; 2011; 36: 42–47. Available: http://www.gnu.org/s/parallel

115. Csilléry K, Blum MGB, Gaggiotti OE, François O, Csillery K, Francois O. Approximate Bayesian Computation (ABC) in practice. Trends Ecol Evol. 2010; 25: 410–418. https://doi.org/10.1016/j.tree.2010.04.001 PMID: 20488578

116. Bertorelle G, Benazzo A, Mona S. ABC as a flexible framework to estimate demography over space and time: some cons, many pros. Mol Ecol. 2010; 19: 2609–2625. https://doi.org/10.1111/j.1365-294X.2010.04690.x PMID: 20561199

117. Csilléry K, François O, Blum MGB. abc: an R package for approximate Bayesian computation (ABC). Methods Ecol Evol. 2012; 3: 475–479. https://doi.org/10.1111/j.2041-210X.2011.00179.x

118. Blum MGB, François O. Non-linear regression models for Approximate Bayesian Computation. Stat Comput. 2010; 20: 63–73. https://doi.org/10.1007/s11222-009-9116-0

119. Weir B, Cockerham C. Estimating F-Statistics for the Analysis of Population Structure. Evolution (N Y). 1984; 38: 1358–1370. Available: http://www.jstor.org/stable/2408641

120. Danecek P, Auton A, Abecasis G, Albers C a, Banks E, DePristo M a, et al. The variant call format and VCFtools. Bioinformatics. 2011; 27: 2156–8. https://doi.org/10.1093/bioinformatics/btr330 PMID: 21653522

121. Smadja CM, Canbäck B, Vitalis R, Gautier M, Ferrari J, Zhou J-J, et al. Large-scale candidate gene scan reveals the role of chemoreceptor genes in host plant specialization and speciation in the pea aphid. Evolution. 2012; 66: 2723–38. https://doi.org/10.1111/j.1558-5646.2012.01612.x PMID: 22946799

122. Harte D. HiddenMarkov: Hidden Markov Models. R package version 1.8–7. Wellington: Statistics Research Associates; 2016.

123. Catchen J, Hohenlohe P a., Bassham S, Amores A, Cresko W a. Stacks: an analysis tool set for population genomics. Mol Ecol. 2013; 22: 3124–3140. https://doi.org/10.1111/mec.12354 PMID: 23701397

124. Wu TD, Nacu S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics. 2010; 26: 873–81. https://doi.org/10.1093/bioinformatics/btq057 PMID: 20147302

125. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009; 25: 2078–9. https://doi.org/10.1093/bioinformatics/btp352 PMID: 19505943

126. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics. 2000; 155: 945–959. PMID: 10835412

127. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: A tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007; 81: 559–575. https://doi.org/10.1086/519795 PMID: 17701901

128. Jombart T, Ahmed I. adegenet 1. 3–1: new tools for the analysis of genome-wide SNP data. Bioinformatics. 2011; 27: 3070–3071. https://doi.org/10.1093/bioinformatics/btr521 PMID: 21926124

129. Earl D a., vonHoldt BM. STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. Conserv Genet Resour. 2011; 4: 359–361. https://doi.org/10.1007/s12686-011-9548-7

130. Gompert Z, Alex Buerkle C. Introgress: A software package for mapping components of isolation in hybrids. Mol Ecol Resour. 2010; 10: 378–384. https://doi.org/10.1111/j.1755-0998.2009.02733.x PMID: 21565033

131. Larson EL, Andrés JA, Bogdanowicz SM, Harrison RG. Differential introgression in a mosaic hybrid zone reveals candidate barrier genes. Evolution (N Y). 2013; 67: 3653–3661. https://doi.org/10.1111/evo.12205 PMID: 24299416

132. Excoffier L, Dupanloup I, Huerta-Sánchez E, Sousa VC, Foll M. Robust demographic inference from genomic and SNP data. PLoS Genet. 2013; 9: e1003905. https://doi.org/10.1371/journal.pgen.1003905 PMID: 24204310

133. Bindea G, Mlecnik B, Hackl H, Charoentong P, Tosolini M, Kirilovsky A, et al. ClueGO: A Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. Bioinformatics. 2009; 25: 1091–1093. https://doi.org/10.1093/bioinformatics/btp101 PMID: 19237447

134. Christmas Rowan; Avila-Campillo Iliana; Bolouri Hamid; Schwikowski Benno; Anderson Mark; Kelley Ryan; Landys Nerius; Workman Chris; Ideker Trey; Cerami Ethan; Sheridan Rob; Bader Gary D.; Sander C. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Am Assoc Cancer Res Educ B. 2005; 12–16. https://doi.org/10.1101/gr.1239303.metabolite

135. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B. 1995; 57: 289–300. Available: http://www.jstor.org/stable/10.2307/2346101