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

Short title: High divergence with localized introgression during speciation

Key words: stickleback, migration, speciation continuum, recombination, neo-sex chromosome

Abstract

Speciation is a continuous process and analysis of species pairs at different stages of divergence provides insight into how it unfolds. Genomic studies on young species pairs have often revealed peaks of divergence and heterogeneous genomic differentiation. Yet it remains unclear how localised peaks of differentiation progress to genome-wide divergence during the later stages of speciation with 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 known of the genomic signatures of divergence and introgression in older 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 hybridise at a low rate throughout divergence. Population genomic data revealed that high levels of genomic differentiation are maintained across the majority of the genome when gene flow occurs. However despite this, we identified multiple, small regions of introgression, strongly correlated with 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.

Author summary

When species evolve, reproductive isolation leads to a build-up of differentiation in the genome where genes involved in the process occur. Much of our understanding of this comes from early stage speciation, 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 [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] with some exceptions [9,13]. Even for these cases [13,14], it is unclear whether gene flow has occurred throughout divergence or whether the species pair in question experienced periods of geographical isolation (see below). A bias towards early stage divergence has undoubtedly been informative to understand the onset of the speciation process [15–18]. However, it is limited in its scope for understanding the factors that eventually lead to the completion of the speciation process – i.e. the evolution of genome-wide differentiation. When divergence with gene flow is studied across the speciation continuum in *Timema* stick insects (including the later stage divergence) the range of genome-wide differentiation appears to be disjointed with a gap in F_{ST} between 0.3 and 0.7 [8]. Furthermore, literature surveys have also indicated that many species pairs exhibit a mean F_{ST} between either 0.03-0.27 or 0.70-0.90 when gene flow is occurring [8,9]. This disjointed distribution of differentiation is consistent with the idea that progression towards speciation may be non-linear, with a phase transition due to genome wide congealing [10,19].

100 In addition to need for more balanced understanding of the extent of
101 genomic divergence and introgression at later stages of the speciation
102 continuum, there is a need for studies which account for factors such as
103 demographic history [12,14]. This is because high genome-wide differentiation
104 in a single species pair may have evolved via genetic drift and local adaptation
105 during allopatric isolation, rather than due to divergence with gene flow. In other
106 cases, heterogeneous genomic differentiation may be due to erosion of genetic
107 differentiation due to introgression following secondary contact after
108 geographical isolation. Without a picture of the demographic history, this
109 scenario may be indistinguishable from primary divergence [20] and will
110 therefore introduce bias if multiple species pairs are compared. Despite the fact
111 that the expected pattern of genomic differentiation during speciation is
112 influenced by the timing and duration of geographical isolation [7], testing
113 different demographic histories has been somewhat neglected by the field [7,20].
114 Other factors besides demographic history of a species pair can also
115 confound patterns of heterogeneous genomic differentiation. For example, local
116 adaptation or background selection in genomic regions where recombination is
117 reduced can elevate differentiation measures and be mistaken for barrier loci
118 [13,21,22]. Similarly, regions of low differentiation may be caused by not only
119 on-going gene flow but also incomplete lineage sorting (ILS) [21,23] and
120 mutation rate variation [24]. Distinction between gene flow and ILS is likely
121 easier in more divergent species pairs [24–26]. Furthermore, the use of multiple
122 classical and recently developed methods, such as detection of recent hybrid
123 progeny, ABBA-BABA tests [27,28], model-based inference[29], and comparison
124 between allopatric and sympatric pairs [22,28] provide a means to distinguish

125 signatures of gene flow from alternative explanations. Thus, there is a need for
126 studies which account for factors such as demographic history, recombination
127 rate variation, and ILS that can confound the interpretation of genome scan data
128 [7,12].

129 Three-spined stickleback species pairs (genus *Gasterosteus*) span the
130 speciation continuum at varying stages of divergence, making them a model
131 system for speciation research [30,31]. To-date genomic research on speciation
132 with gene flow in the stickleback complex has largely focused on weakly
133 divergent species pairs, such as lake-stream ecotypes, with mean genome-wide
134 F_{ST} values of less than 0.3 [32–34]. Such studies have shown that the genomic
135 landscape of differentiation between these recently diverged sympatric or
136 parapatric species pairs is heterogeneous and interspersed with multiple peaks
137 of high differentiation [16,32,34]. The emerging pattern is consistent with
138 predictions under the genic concept of speciation – i.e. that reproductive
139 isolation is localized in the genome at early stages of speciation [4,35]. However,
140 it remains unclear whether such localized differentiation will eventually
141 progress toward genome-wide differentiation in the face of gene flow.

142 Toward the end of the stickleback speciation continuum is a marine
143 species pair in Japan [36,37]. The Japan Sea stickleback (*G. nipponicus*) is
144 sympatric with the Pacific Ocean lineage of three-spined stickleback (*G.*
145 *aculeatus*) (Fig 1A) in the waters surrounding the Japanese archipelago (Fig 1C)
146 [36,38]. Divergence time between the two marine species has been estimated to
147 be 1.5-2 million years based on allozyme and microsatellite data [37,39], making
148 it much older than postglacial stickleback species pairs. Speciation has been
149 hypothesized to have occurred as a result of the repeated isolation of the Sea of

150 Japan during the Pleistocene, but this divergence scenario remains to be
151 explicitly tested [37,39]. A unique feature of the *G. nipponicus* and *G. aculeatus*
152 system, relative to postglacial stickleback species pairs, is that a neo-sex
153 chromosome has arisen due to a fusion between a Y chromosome and a
154 previously autosomal chromosome IX (chrIX) in the *G. nipponicus* lineage
155 [36,40]. Furthermore, crosses between Japan Sea females and Pacific Ocean
156 males show hybrid male sterility [37]. Previous quantitative trait locus (QTL)
157 mapping identified QTL for courtship behaviour on the neo-X and hybrid male
158 sterility on the ancestral-X. However, there are other isolating barriers, such as
159 eco-geographical isolation, temporal isolation, and ecological selection against
160 migrants [37,41,42]. The combination of these multiple barriers most likely
161 contributes to the strong reproductive isolation in this system [36,43]. However,
162 despite such strong divergence, hybrids have been observed where the two
163 species co-occur in Northern Japan [36] and mitochondrial discordance between
164 the species suggests some history of introgression during speciation [44,45].
165 Although the Japanese species pair therefore represents the furthest point of
166 divergence along the stickleback speciation continuum, speciation remains
167 incomplete. The evolutionary history and genome-wide patterns of genetic
168 differentiation and introgression of this strongly divergent species pair therefore
169 remains an open question

170 The aim of our study was to address gap in our knowledge; i.e. to quantify
171 the patterns of genomic differentiation and introgression at a later stage of the
172 speciation continuum. To this end, we used a whole-genome reseqeuncing data
173 from the Japanese stickleback species pair to determine their evolutionary
174 history and characterise patterns of gene flow between them. Our first aim was

175 to establish how and when divergence took place between *G. nipponicus* and *G.*
176 *aculeatus*. Using thousands of genomic loci and a coalescent modelling approach,
177 we tested a range of divergence scenarios and estimated the timing and duration
178 of isolation, the extent of gene flow and fluctuations in population size. After
179 identifying that the two species have indeed diverged in the face of gene flow, we
180 then used a comparative genome scan approach with an additional *G. aculeatus*
181 lineage from the Atlantic Ocean [46] as an allopatric control (Fig 1A, Fig S1).
182 After establishing that gene flow has occurred but that a high level of genomic
183 differentiation has remained, we used two independent measures of gene flow to
184 identify where in the genome introgression has left its mark. We tested whether
185 introgression occurs more frequently in regions of high recombination and
186 whether it occurs in regions with functionally important genes. Our findings
187 suggest a high level of genome-wide divergence can be maintained, as
188 introgression is restricted to small, localized genomic regions.
189

190 ## Results

191 **Ancestral demography and population genomic analyses support**
192 **divergence with gene flow**

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

199 We calculated genealogical sorting index (*gsi*) [47] on maximum
200 likelihood phylogenies estimated from non-overlapping sliding windows of 20
201 kb across the autosome. High *gsi* indicates monophyly, while low *gsi* indicates
202 mixed ancestry [47]. Genome-wide averages (± SD) of *gsi* were high, but not
203 complete, for all three *Gasterosteus* lineages with that of the Japan Sea
204 stickleback being the highest (Atlantic *gsi* = 0.45 ± 0.10, Pacific *gsi* = 0.57 ± 0.09,
205 Japan Sea *gsi* = 0.72 ± 0.06).
206 This is in stark contrast to the mitogenome phylogeny where
207 sticklebacks from both species occurring in Japan fall into a single clade separate
208 from the clade occurring in the Western Pacific and Atlantic (Fig 1B, Fig S2). A
209 lack of mitogenome divergence between *G. aculeatus* and *G. nipponicus* from the
210 Japanese archipelago suggests mitochondrial introgression has occurred where
211 these lineages overlap (Fig 1C).
212 Since the consensus autosomal phylogeny suggests a more recent split
213 between the Pacific and Atlantic *G. aculeatus* lineages, a deeper mtDNA split
214 based on spatial distribution suggests mitochondrial introgression has likely
215 occurred from the Japan Sea *G. nipponicus* into the Pacific Ocean *G. aculeatus*.
216 Divergence time estimates between the mitogenome clades are thus informative
217 for dating speciation. Bayesian coalescent analysis using a strict clock model in
218 Bayesian Evolutionary Analysis by Sampling Trees (BEAST) suggests a median
219 split date of 1.30 million years (0.15-2.41; 95% Highest Posterior Density [HPD]
220 intervals; Table S2) for the two major mitogenome clades (Fig S2), consistent
221 with previous estimates [44]. Divergence between Eastern Pacific and Atlantic
222 haplotypes is more recent at 0.39 million years (0.03-0.74; 95% HPD) but is
223 older than the Most Recent Common Ancestor (MRCA) of all haplotypes

224 occurring in Japan (Fig 1B, S2A), suggesting mitochondrial gene flow from *G.*
225 *nipponicus* to *G. aculeatus* has occurred in the recent past (i.e. <0.39 million years
226 BP).

227 Mitochondrial introgression can be driven by large demographic
228 disparities between populations with gene flow and is more likely to occur from
229 a larger to a smaller population [48]. To investigate the demographic disparities
230 between *G. aculeatus* and *G. nipponicus*, we used pairwise sequential Markov
231 coalescent (PSMC) on all 26 Atlantic Ocean, Japan Sea and Pacific Ocean
232 resequenced stickleback genomes. Strikingly, *G. nipponicus* experienced a severe
233 bottleneck around 0.15-0.3 million years before present (BP) (Fig 1D); mean N_e
234 fell to 26,422 ± 1,191 at its lowest point. Subsequently after 0.1 million years BP,
235 *G. nipponicus* underwent a dramatic population size expansion (Fig 1D): mean N_e
236 rose to 195,974 ± 28,832 (i.e. ~7.5 times increase from the bottleneck) during
237 the late Pleistocene. In contrast, the Japanese Pacific Ocean *G. aculeatus*
238 population has remained relatively stable throughout its history (mean N_e ± SD =
239 118,150 ± 4,330; Fig 1D, see Fig S3 for bootstrap support). Although the Atlantic
240 (Fig 1D) and Western Pacific lineages of *G. aculeatus* (Fig S4) also experienced
241 some population growth during the late Pleistocene, their effective population
242 sizes remained smaller than that of *G. nipponicus*. Genome-wide averages of
243 Tajima's *D* also support a recent demographic expansion for *G. nipponicus* (mean
244 ± SD of Tajima's *D* = -0.82±0.45) and stable population size in the Pacific Ocean
245 (mean ± SD of Tajima's *D* = -0.04 ± 0.63). Taken together, these data indicate that
246 mitochondrial introgression likely occurred during the late Pleistocene, when *G.*
247 *nipponicus* N_e was substantially larger than *G. aculeatus* N_e. This is consistent

248 with the hypothesis that mitochondrial gene flow from *G. nipponicus* to *G.*
249 *aculeatus* has occurred <0.39 million years BP (see above).
250 To explicitly test whether divergence between *G. aculeatus* and *G.*
251 *nipponicus* occurred in the presence of gene flow, we used an Approximate
252 Bayesian Computation (ABC) approach with 1,874 2 kb loci randomly sampled
253 from across the autosome. We tested five divergence scenarios – isolation (I),
254 isolation with migration (IM), isolation-with-ancient-migration (IAM), isolation-
255 with-recent-migration (IRM) and isolation-with-ancient-and-recent-migration
256 (IARM) – i.e. two discrete periods of contact. Since the results of our PSMC
257 analyses clearly indicate N_e has varied throughout divergence (Fig 1D), we
258 performed a hierarchical ABC analysis, first selecting the most appropriate
259 population growth model (i.e. constant size, population growth and a Japan Sea
260 bottleneck) within each divergence scenario and then performing final model
261 selection amongst the best supported divergence/growth model scenarios (see
262 Supplementary Methods for full specification of models, priors, parameters and
263 extensive sensitivity testing).
264 Using 20 summary statistics (see Supplementary Methods for a full list
265 of statistics used) and a neural-network rejection method with 1% tolerance of
266 simulated datasets, the best-supported divergence scenario was a model of IM
267 (Fig 2A; Table 1). Parameter estimates from the IM model suggest divergence
268 between *G. aculeatus* and *G. nipponicus* occurred 0.68 million years ago (median
269 estimate, 0.18-4.17 million years, lower & upper 95% HPD; Fig 2B). A Japan Sea
270 bottleneck occurred 0.3 million years ago (0.03-2.21 million years 95% HPD),
271 reducing N_e to about 20% of the contemporary estimate (Fig 2C, Table S3). Mean
272 migration rates between the two species were low, and migration from the Japan

273 Sea into the Pacific Ocean lineage was slightly greater (Fig 2D, Table S3).
274 Contemporary N_e of the Japan Sea lineage is larger than that of the Pacific Ocean,
275 although the N_e estimates differed in magnitude from those estimated by PSMC
276 (Figs 1D and 2C, Table S3).

277 Identifying admixture between species where they co-occur provides
278 strong evidence of on-going introgression [7,12]. To address this, we used a
279 RAD-sequencing dataset with a larger sample size of 245 individuals from the
280 Atlantic, Pacific and Japan Sea lineages, including previously published data from
281 Pacific -derived populations in North America [49]. Principal component analysis
282 (PCA) of allele frequencies at 3, 744 high-quality bi-allelic SNPs showed that,
283 consistent with our whole genome data, the main axis explaining 20% of the
284 variance was between *G. aculeatus* and *G. nipponicus* (Fig S5). The secondary axis
285 explaining 9.49% of the variance was mainly between the Atlantic and Pacific
286 populations (Fig S5). Importantly, PCA shows a single individual is intermediate
287 between the Pacific and Japan Sea populations occurring in Akkeshi, the
288 sympatric site in Hokkaido, Japan where our whole genome-sequenced samples
289 were collected (Fig 1C). A separate Bayesian analysis for admixture using
290 STRUCTURE [50,51] found greatest support for $K = 2$ in the Japanese populations
291 and also identified the putative F_1 hybrid plus individuals with possible recent
292 admixture at this sympatric site (Fig S6).

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

298

299 **High levels of genome-wide divergence with highly localized signatures of**

300 **introgression**

301 Genome-wide differentiation was strikingly high between *G. nipponicus* and *G.*

302 *aculeatus* regardless of their geographical overlap: both relative (F_{ST}) and

303 absolute divergence (d_{XY}) were high (Fig 3A & B, Fig 4, and Figs S7 and S8). The

304 genome-wide average of F_{ST} between the sympatric species was 0.628; this is

305 higher than all other studied stickleback species pairs [32–34,52] (see Fig 3C).

306 Despite consistently high divergence, both F_{ST} and d_{XY} values were significantly

307 lower where the two species occur in contact (Table 2, Figs 3A & B, 4, S7 and S8;

308 10,000 replicate permutation tests on 10 kb windows: $P < 2.2 \times 10^{-16}$ for both

309 statistics), consistent with the presence of gene flow in sympatry.

310 A more fine-scale analysis of genome-wide divergence based on 10 kb

311 non-overlapping windows revealed that the high baseline divergence between *G.*

312 *nipponicus* and *G. aculeatus* is interspersed by regions of low differentiation in

313 both F_{ST} and d_{XY} genome scans (Figs 4 top two panels, S7 and S8). Processes such

314 as background selection that alter within-population diversity can bias relative

315 measures of differentiation such as F_{ST} [7,21,22,53]. Furthermore, absolute

316 measures like d_{XY} can take a long time to reach equilibrium between populations

317 following divergence, meaning a lower power for detecting recent introgression

318 events [7,26]. Therefore, to better identify genomic regions of recent

319 introgression, we calculated two independent measures of introgression. The

320 first of these was G_{MIN}, the ratio of the minimum d_{XY} to the average d_{XY} [26].

321 Under strict isolation, G_{MIN} is the lower bound of divergence time between two

322 populations, whereas when introgression occurs, G_{MIN} reflects the timing of the

323 most recent migration [26]. The second measure was f_d, an estimate of the
324 proportion of introgressed sites in a genome window, calculated using a four
325 population ABBA-BABA test [54]. G_{MIN} is more effective at identifying recent, low
326 level gene flow than either F_{ST} or d_{XY} but by definition it is unable to detect
327 genomic regions where complete introgression has occurred [26], which can
328 however be detected using f_d. Importantly, both measures are robust to variation
329 in recombination rate [26,54]. Combining these two statistics therefore allows us
330 to identify both low-level (G_{MIN}) and strong introgression (f_d).

331 Focusing on between species comparisons, mean (± SD) G_{MIN} measured
332 from 10 kb non-overlapping windows was greater in allopatry than sympatry
333 (Japan Sea vs. Atlantic: 0.876 ± 0.071; Japan Sea vs. Pacific: 0.857±0.103;
334 randomization test $P < 2.2\text{x}10^{-16}$; Fig 4). Mean f_d was also greater when the
335 species overlapped (JS vs. AT: -0.0031 ± 0.0540; JS vs. PO: 0.0039±0.0328; $P <$
336 $2.2\text{x}10^{-16}$; Fig 4), and both statistics are more strongly negatively correlated in
337 sympatry (Fig S9) supporting gene flow between *G. nipponicus* and Japanese
338 populations of *G. aculeatus*.

339 Genomic regions of low G_{MIN} (i.e. G_{MIN} valleys) may indicate recent
340 introgression. We identified genome windows with low G_{MIN} values using a
341 Hidden-Markov classification model [55] and an outlier approach based on
342 permutations of variant sites (Fig S10). We then clustered 10 kb outlier
343 windows occurring within 30 kb of one another into putative G_{MIN} valleys. G_{MIN} in
344 particular may be susceptible to false positives as a result of ILS. However, lower
345 d_{XY} and higher f_d in sympatric G_{MIN} valley windows compared to the genomic
346 background suggests ILS alone does not explain the patterns observed here (Fig
347 S11; randomization test, $P < 2.2 \text{ x } 10^{-16}$ in both cases). These regions of

348 introgression were more common in the genome when the two species
349 overlapped, with 637 valleys in sympatry (JS-PO comparison) compared to 337
350 in allopatry (JS-AT comparison) (randomization test, t = 5.35, $P < 2.2 \times 10^{-16}$) and
351 a greater number of valleys per chromosome (Fig 5A), although mean valley size
352 did not differ significantly (77.6 kb and 75.4 kb in sympatry and allopatry
353 respectively, $P = 0.82$). Interestingly, 225 valleys were shared between JS-PO and
354 JS-AT comparisons (Fig. 4). These shared valleys may indicate ILS but they may
355 also reflect introgression from Pacific Ocean to Japan Sea, where one or a few
356 Japan Sea individuals carry haplotypes derived from Pacific Ocean and therefore
357 are also similar to Atlantic Ocean haplotypes too. However, a larger number of
358 valleys (412 valleys) were unique to the JS-PO comparison, where introgression
359 might occur from Japan Sea to Pacific Ocean.

360 A similar geographical comparison of peaks of f_d between species was
361 not possible, due to the much lower genome-wide distribution of f_d between *G.*
362 *nipponicus* and the Atlantic *G. aculeatus* (Fig 4). Nonetheless, Hidden-Markov
363 classification identified 823 f_d peaks occurring between *G. nipponicus* and Pacific
364 *G. aculeatus* (Fig S12). If the f_d peaks are mainly indicate introgression from
365 Pacific Ocean to Japan Sea, d_{XY} between Japan Sea and Atlantic Ocean is expected
366 to be lower in these regions compared to the genome background, as Japan Sea
367 fish carry haplotypes derived from the Pacific Ocean, which in turn are similar to
368 the Atlantic Ocean haplotypes. While JS-AT d_{XY} was lower in f_d peaks compared to
369 the genome background (one-tailed permutation test, $P < 2.2 \times 10^{-16}$), this
370 difference was not very clear (Fig. S13). In contrast, if introgression occurred
371 mainly from Japan Sea to Pacific Ocean, d_{XY} in the PO-AT comparison should
372 increase in f_d peaks relative to the genome background, as Pacific Ocean fish

373 carry Japan Sea-derived haplotypes, which are divergent from the Atlantic Ocean
374 haplotypes. We clearly observed this pattern ($P < 2.2 \times 10^{-16}$; Fig. S13);
375 suggesting that introgression from Japan Sea to Pacific Ocean may be more
376 predominant than the opposite direction.

377 To further investigate the direction of gene flow, we used partitioned *D*
378 statistics (an extension of the four population test – see Fig. S14), which tests the
379 excess of shared derived alleles using five, rather than four populations [56]. To
380 this end, we added an allopatric Japan Sea population (collected from Lake Shinji,
381 a brackish lake at the Japan Sea coast of southern Honshu). A positive D_{12}
382 statistic is proposed to indicate the predominance of introgression from P3 to P2
383 (Fig. S14) [56]. When P3 was set to Japan Sea (where $P3_1$ is sympatric and $P3_2$ is
384 allopatric with the Pacific Ocean) and P2 to Pacific Ocean (see Fig. S14B), D_{12} was
385 significantly positive in f_d peaks (one-tailed permutation test, $P < 2.2 \times 10^{-16}$). In
386 contrast, when we rotated the populations at the tips – i.e. setting P2 to
387 sympatric Japan Sea, $P3_1$ to Pacific Ocean, and $P3_2$ to Atlantic Ocean (see Fig
388 S14C), D_{12} was not positive, consistent with the suggestion that introgression is
389 occurring mainly from Japan Sea to Pacific Ocean. However, the resolution of
390 partitioned *D* statistics has been criticized [57]; positive D_{12} can also be caused
391 by introgression from the Pacific Ocean (P2) to the common ancestor of the
392 sympatric and allopatric Japan Sea populations ($P3_1$ & $P3_2$). To overcome this
393 issue, we calculated D_{FOIL}, which also uses a five-population test but accounts for
394 all possible introgression events [57]. When P_1 = sympatric Japan Sea, P_2 =
395 allopatric Japan Sea, P_3 = Pacific Ocean, and P_4 = Atlantic Ocean (Fig. S15), D_{FOIL}
396 clearly indicated the presence of ancestral introgression (239 out of 4,236 100
397 kb-windows) between the Japan Sea ancestor (P_{12}) and the Pacific Ocean (P_3)

398 (see Fig. S15). However, we found only a few windows showing unidirectional
399 introgression (6 loci in total), precluding conclusion of the predominant
400 direction of introgression using this analysis (Fig. S15). This low sensitivity may
401 be due to the fact that structuring in the Japan Sea lineage is low [58] – i.e. recent
402 divergence time or high intraspecific gene flow.
403
404 **Characterization of genomic regions of introgression**
405 To investigate whether introgression co-varies with recombination rate, we used
406 a previously published recombination map from an Atlantic *G. aculeatus* cross
407 [59] to interpolate genome-wide recombination rate variation (see Methods).
408 We detected a negative correlation between recombination rate and G_{MIN} and a
409 positive correlation with f_d (Pearson's correlation, G_{MIN}: r = -0.17, $P < 2.2 \times 10^{-16}$;
410 f_d: r = -0.08, $P < 2.2 \times 10^{-16}$, Fig S16). Accordingly, mean recombination rate for
411 putatively introgressed regions was over two times higher than the genome
412 background (G_{MIN}: valley = 8.98 cM/Mb, non-valley = 3.99 cM/Mb; f_d: peak = 9.64
413 cM/Mb, non-peak = 4.16 cM/Mb; randomization test $P < 2.2 \times 10^{-16}$ in both cases;
414 Fig 5B).
415 Sex chromosomes likely played an important role in speciation between
416 *G. aculeatus* and *G. nipponicus*. A fusion between Y and chrIX means that chrIX
417 segregates as a neo-sex chromosome in *G. nipponicus* but not *G. aculeatus* which
418 only carries the ancestral and shared sex chromosome, chrXIX [36,40]. The
419 divergent XY (*G. aculeatus*) and X_1X_2Y (*G. nipponicus*) systems means that
420 recombination is reduced for chrIX and chrXIX in hybrids carrying the neo-Y
421 [40]. Given this recombination rate reduction and previously identified QTL for
422 traits involved in reproductive isolation that map to chrIX and chrXIX [36,40], we

423 tested whether recent introgression (i.e. measured using G_{MIN}) was reduced in
424 this part of the genome relative to the autosome. For this, we repeated our
425 analyses using females only (5 Japan Sea and 6 Pacific Ocean). The number and
426 density of valleys was lowest on the neo-sex chromosome, chrIX (16 valleys or
427 0.8 valleys per Mb) but not the ancestral sex chromosome (chrXIX, see Table S4).
428 Finally, we investigated the nature of introgression between the two
429 species. We first asked whether introgression occurs more frequently in genic or
430 non-genic regions. We identified 3,261 genes occurring in G_{MIN} valleys and 2,958
431 genes from f_d peaks between sympatric *G. aculeatus* and *G. nipponicus*; 60% of
432 genes identified were found in both types of introgressed window, whereas 23%
433 occurred only in G_{MIN} valleys and 15% only in f_d peaks (see Fig S17). Irrespective
434 of the method used to detect putatively introgressed regions, the number of
435 genes identified was greater than the number expected by chance ($P < 0.0001$
436 based on a null distribution generated from 1,000 random samples of the
437 genome). This suggests that introgression may be more likely in genic regions of
438 the genome than non-genic regions.
439 To further investigate the functional enrichment of the genes occurring
440 in regions of introgression, we performed gene ontology (GO) analysis on 2,310
441 G_{MIN} valley and 2,217 f_d peak genes with orthologs in the human genome.
442 Enriched GO terms for f_d peaks included immune response, metabolic processes
443 and chromatin assembly, while enriched GO terms for G_{MIN} valleys included
444 major histocompatibility complex (MHC) protein and metabolic processes (Table
445 S5 & S6).
446

447 **Discussion**

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

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

455 Firstly, contrasting mitochondrial and nuclear genome phylogenies
456 show that mitochondrial introgression has occurred from *G. nipponicus* into *G.*
457 *aculeatus* at some point in the last 0.39 million years. Our mitogenome
458 phylogeny confirmed previous findings that there is no mitochondrial
459 divergence between the *G. nipponicus* and Japanese populations of *G. aculeatus*
460 [44,45]. This is in contrast to our nuclear autosomal phylogeny which showed
461 that majority of the genome supports a clear split between *G. nipponicus* and *G.*
462 *aculeatus* occurring in Japan and that the latter shares a more recent common
463 ancestor with Atlantic European *G. aculeatus* populations. In short, mitogenome
464 data clusters the *Gasterosteus* lineages by geography, while the nuclear data
465 clusters them by species. Introgression is the most likely explanation for this
466 mitonuclear discordance as the lower effective population size of the
467 mitochondrial genome relative to the autosome makes ILS unlikely, particularly
468 over a 1 million year time-scale of divergence.

469 Disparities in effective population size between lineages are a common
470 cause of unidirectional mitonuclear introgression with introgression likely
471 occurring from a larger to a smaller population [48]. Our reconstruction of
472 temporal variation in effective population size using PSMC showed a rapid

473 population explosion *of G. nipponicus* during the late Pleistocene that created a
474 large demographic disparity with the *G. aculeatus* Pacific Ocean lineage, although
475 it should be noted that admixture can increase effective population size
476 estimates when using PSMC [60,61]. Reasons for this population growth remain
477 unclear but it is surprising, particularly since *G. nipponicus* is unable to colonise
478 freshwater environments [42,58], which might be expected to increase effective
479 population size due to meta-population dynamics. A possible explanation is that
480 a lack of dependence on freshwater for spawning [41,62,63] and greater foraging
481 efficacy on marine prey [42] means *G. nipponicus* is better adapted for exploiting
482 an abundant marine environment. Unidirectional mitochondrial introgression
483 might also be caused by female mate choice [64]. Our previous behavioural
484 studies indicate that Japan Sea females often mate with Pacific Ocean males,
485 while Pacific Ocean females rarely mate with Japan Sea males [36,37]. Hybrid
486 females from Japan Sea female and Pacific Ocean male crosses are fertile [37]
487 and will carry Japan Sea mitochondrial DNA. Backcrossing of these hybrids to
488 Pacific Ocean males would result in unidirectional mitochondrial introgression
489 from the Japan Sea to Pacific Ocean.
490 Secondly, our ABC analysis supported a model of isolation with
491 migration. Previously, it has been speculated that the Japan Sea stickleback
492 diverged largely as a result of geographical isolation in the Sea of Japan caused
493 by sea level fluctuation during the early Pleistocene [37,39]. Using ABC, we were
494 able to explicitly test several divergence hypotheses in a statistical framework
495 [65]; our findings suggest that gene flow has likely occurred throughout the
496 majority of divergence history. It should be noted that ABC and most established
497 demographic inference methods perform poorly when resolving the timing of

498 gene flow between lineages [66,67]. Therefore, one caveat to the interpretation
499 of our ABC results is that we cannot rule out the possibility that the two species
500 diverged in repeated cycles of contact (i.e. akin to our IARM model which had the
501 second highest level of support; Table 1), but these periods of contact were
502 simply too close in time. Nonetheless, the pooled posterior probabilities from the
503 analysis overwhelmingly support a model of divergence with gene flow
504 irrespective of the timing or nature of the actual speciation event. The presence
505 of extant recent hybrids in sympatry also strongly indicates that hybridization is
506 still on-going. We observed a probable F_1 hybrid in the wild and several other
507 individuals with evidence of hybrid ancestry in our RAD-seq dataset, consistent
508 with previous studies that observed wild caught hybrids [36,68]. This provides
509 direct observation of admixture in the wild.

510 Finally, lower levels of genome-wide divergence (both F_{ST} and d_{XY})
511 between sympatric pairs compared to allopatric pairs also indicate the presence
512 of gene flow. Our G_{MIN} and f_d genome scans showed a higher number of
513 putatively introgressed regions between *G. nipponicus* and Japanese Pacific *G.*
514 *aculeatus* than between *G. nipponicus* and Atlantic *G. aculeatus*, suggesting that
515 introgression has been occurring even after the Atlantic and Pacific stickleback
516 populations diverged approximately 400,000 years BP. Our partitioned *D*
517 statistics demonstrated that gene flow from *G. nipponicus* into Japanese Pacific *G.*
518 *aculeatus* may be more predominant than the opposite direction in sympatry.
519

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

521 Compared to young species pairs, less is known about the patterns of genomic
522 differentiation at more advanced stages of speciation with gene flow. Our study

523 provides insight on this under-represented stage of divergence. Our ABC
524 analyses placed the estimated divergence time of *G. aculeatus* and *G. nipponicus*
525 at 0.68 million years BP. Similarly, our Bayesian coalescent analysis of
526 mitogenome divergence revealed a 1.3 million year split between the Japanese
527 and Atlantic-Pacific *Gasterosteus* mitochondrial clades. An older divergence time
528 is somewhat expected from the mitochondrial genome, given its fourfold lower
529 effective population size [48]. Because splitting of mitochondrial lineages can
530 occur by geographical structuring without speciation [69,70], our estimate of
531 mitochondrial divergence may reflect more ancient geographical structuring that
532 occurred prior to species divergence. Nonetheless, both mitochondrial and
533 nuclear split estimates suggest that divergence between *G. aculeatus* and *G.*
534 *nipponicus* occurred well before the end of the last glacial period. Therefore the
535 Japanese stickleback system is older than all other previously examined
536 postglacial sympatric or parapatric species pairs, which have typically diverged
537 within the last 20,000 years [30].

538 The Japanese stickleback system also has a mean genome-wide F_{ST} value
539 of 0.628, higher than any other sympatric or parapatric stickleback species pair
540 studied so far (Fig 3C); placing this pair at the furthest end of the speciation
541 continuum. A recent meta-analysis quantifying genetic differentiation pointed to
542 a lack of evidence of sympatric or parapatric species pairs with F_{ST} between 0.3
543 and 0.7[8]. The Japanese species pair is therefore an example of an
544 underepresented sympatric species pair that maintains high divergence in the
545 face of gene flow. Within the stickleback species complex, this high divergence is
546 in stark contrast to the typical low baseline divergence interspersed with regions

547 of high differentiation observed in more commonly studied species pairs such as
548 lake-stream or freshwater-anadromous pairs [16,33,71].
549 The primary explanation for the observed elevated divergence is most
550 likely the more ancient divergence time of the Japan Sea-Pacific Ocean species
551 pair compared to postglacial species pairs [16,72]. However, the results of our
552 demographic analyses indicate that high divergence is not due to a long period of
553 allopatric isolation without gene flow, contrary to what has previously been
554 suggested [37,39]. This is important, as failing to account for variation in
555 evolutionary history among species pairs placed on a continuum will obscure the
556 processes leading to higher differentiation as speciation progresses. A further
557 explanation for the high genomic divergence is the presence of strong isolating
558 barriers between the Japan Sea and Pacific Ocean sticklebacks. Total
559 reproductive isolation (0.970) is greater than in all postglacial species pairs
560 (0.716-0.895) [43] and arises from a combination of habitat [41,42], temporal
561 [62] and sexual isolation, and hybrid sterility [36,37]. Recent theoretical studies
562 have shown that selection on many barrier loci in the face of gene flow may
563 result in a transition from low to high differentiation as a result of 'genome-wide
564 congealing' [10,19]. It is important to note that we lack evidence that such a
565 transition might explain the high differentiation we see here relative to the rest
566 of the stickleback continuum (Fig 3C). However, pervasive selection on multiple
567 isolating barriers in the Japanese stickleback system does suggest that this
568 process could have contributed to the high genomic differentiation we observe
569 [5].
570
571 **Localized introgression at a late stage of speciation with gene flow**

572 Our study has also demonstrated two important signatures of introgression in
573 the Japanese sympatric stickleback pair. Firstly, levels of background genome
574 differentiation between *G. aculeatus* and *G. nipponicus* estimated by F_{ST} were
575 lower in sympatry compared to allopatry. The higher overall genetic
576 differentiation between *G. nipponicus* and Atlantic *G. aculeatus* is likely due to
577 genetic drift and local adaptation and the fact that these two lineages have never
578 overlapped geographically. Secondly and strikingly, we identified small regions
579 of localised introgression dispersed throughout the genome when *G. nipponicus*
580 and *G. aculeatus* co-occur in sympatry. These introgression regions were
581 measured using G_{MIN}, the ratio of minimum d_{XY} to mean d_{XY}, where low values
582 indicate introgression [26] and f_d, the proportion of introgressed sites in a
583 genome window [54].

584 Several methodological issues might influence these measures of
585 introgression. Firstly, both G_{MIN} and f_d are sensitive to sample size; fewer
586 individuals will mean rare haplotypes have a lower sampling probability.
587 However by re-conducting our analyses using only females, a much smaller
588 sample size than our main analysis, we still identified clear signals of
589 introgression. Secondly, G_{MIN} will be biased downwards if a recently backcrossed
590 individual is included in the dataset. All Japanese *G. aculeatus* and *G. nipponicus*
591 used in the study were identified as 'pure' individuals with genotyping at
592 multiple microsatellite loci prior to resequencing [36,40]. Furthermore, we
593 examined the identity of the two haplotypes producing the lowest value of d_{XY} in
594 G_{MIN} valleys and confirmed that these came from different individuals in each
595 case (data not shown). Nonetheless, even a single haplotype segregating in a
596 population due to ILS can lower the value of G_{MIN}; i.e. ILS could produce small,

597 interspersed G_{MIN} valleys. However, ILS cannot explain why more introgressed
598 regions occur in sympatry (Fig 5A) or why there is a clear association with an
599 increased proportion of introgressed sites (i.e. f_{d}) and G_{MIN} (Fig S9 & S13).
600 What then underlies the localised pattern of introgression we observe?
601 One possible explanation is the fact that many isolating barriers are involved in
602 reproductive isolation [36,43]. Although the genomic basis of these isolating
603 barriers remains unknown, it is likely that barrier loci occur throughout the
604 genome; pervasive selection at multiple loci is expected to limit the extent of
605 introgression at this scale [73]. The strength and extent of selection at a barrier
606 locus and genomic regions linked to it is proportional to recombination rate [73].
607 Recombination determines effective migration rate [74]; when recombination is
608 high, neutral and adaptive loci linked to the target of negative selection in the
609 recipient population have a greater probability of escaping removal and so their
610 probability of introgression is greater [3]. Selection has a higher efficiency in
611 these high recombination rate regions due to increased effective population size
612 – therefore deleterious introgression is also more likely to be removed. The
613 expectation then is that signatures of introgressed neutral or adaptive alleles are
614 most likely to persist in regions of the genome where recombination rate is
615 sufficiently high enough, and indeed, the positive association between
616 introgressed regions and recombination rate we observed supports this (Fig 5B,
617 Fig S15). Introgression is typically lower on sex chromosomes relative to
618 autosomes in multiple taxa due to the effects of reduced recombination and
619 greater exposure to selection in the hemizygous sex [75]. The sex chromosomes
620 play an important role in the Japanese stickleback system, harbouring QTL for
621 hybrid sterility and behavioural isolation [36]. Consistent with this, we observed

622 lower introgression on the neo-sex chromosome (Fig 5E & F), although we
623 cannot exclude the possibility that the fusion occurred more recently than the
624 speciation event, so the opportunity for introgression on the neo-sex
625 chromosomes was simply low relative to the rest of the genome. Taken together,
626 our findings suggest that strong divergent selection and recombination rate
627 variation may determine the localised signature of introgression in the genome.
628 The nature of gene flow in the Japanese stickleback system may also
629 give some clues as to why we observe such highly localised introgression. One
630 possibility is that a proportion of the introgression we detected is adaptive; i.e. it
631 is maintained because of either directional or balancing selection. Adaptive
632 introgression has been detected in a wide range of taxa [76], including humans
633 [77]. However, the expected signatures of the process remain unclear –
634 especially when introgression is widespread in the genome, as is possibly the
635 case here. Our GO analyses suggest an enrichment of immune response genes,
636 including MHC genes, and metabolism genes in introgressed regions. Immune
637 genes have been identified as being under balancing selection between
638 hybridising taxa, particularly plants [78] and birds [79]. Several genes involved
639 in metabolism are also reported to be under balancing selection in humans [80].
640 Furthermore, recent analysis suggests that negative frequency dependent
641 selection might result in introgression of rare MHC alleles between divergent
642 stickleback ecotypes [81]. Further research is necessary to directly test whether
643 this process might explain introgression in the Japanese stickleback system.
644
645 **Conclusion**

646 Much of our knowledge of how genomic differentiation builds along the
647 speciation continuum is drawn from studies focusing on young, allopatric or
648 completely reproductively isolated species pairs. Very few examples of species
649 pairs at a later stage of divergence with on-going gene flow have been
650 investigated. Here, we have shown that the Japan Sea and Pacific Ocean species
651 pair exemplifies this under-represented stage of speciation and is situated at the
652 further end of the stickleback species continuum. The high genomic
653 differentiation between the species may be due to a more ancient divergence
654 time than previously studied postglacial species pairs, selection on multiple
655 isolating barriers or a combination of the two. Despite high differentiation, gene
656 flow is on-going between the species and we identified localized signatures of
657 introgression throughout the genome. Although the localized nature of the
658 introgression remains unclear, selection – either directional or balancing – may
659 play some role in promoting it. Overall, our study demonstrates that high levels
660 of genomic divergence can be established and maintained in the presence of
661 gene flow. Further genomic studies on more species pairs at late stages of
662 speciation with gene flow will help to understand the generality of the patterns
663 seen here.
664

665 ## Materials and Methods

666 ### Ethics Statement

667 All animal experiments were approved by the institutional animal care and use
668 committee of the National Institute of Genetics (23-15, 24-15, 25-18).
669

670 **Sample collection, whole genome resequencing and RAD sequencing**.
671 Collection and sequencing of all Japanese individuals used for whole genome
672 resequencing has been described previously [40] except the allopatric Japan Sea
673 fish. Briefly sympatric populations were captured from the Akkeshi system in
674 Hokkaido, Japan in 2006 (Fig 1C). The allopatric Japan Sea female was collected
675 in Lake Shinji in March 2014. The outgroup species, *G. wheatlandi* was captured
676 from Demarest Lloyd State Park, MA, USA in 2007, as described previously [40].
677 Libraries were constructed with TruSeq DNA Sample Preparation Kit (Illumina)
678 and whole-genome 100 bp paired-end sequencing was performed on an Illumina
679 HiSeq2000 at the National Institute of Genetics (sympatric JS and PO) and
680 Functional Genomics Facility, NIBB Core Research Facilities (allopatric JS) [40].
681 Whole genome sequencing of North American marine and stream populations
682 collected from Little Campbell River, BC, Canada was reported previously
683 [52,82]. For the six Atlantic *G. aculeatus* individuals (North Sea) included in the
684 study, we used previously published sequences [46].
685 Japanese individuals used for RAD sequencing have been previously
686 described elsewhere [58]. Samples used for RAD sequencing from the Atlantic
687 lineage were collected from across Ireland in 2009-2011 [70,83]. DNA was
688 extracted using a Qiagen DNeasy Blood and Tissue Kit (Qiagen, Valencia, CA,
689 USA). Single digest RAD-sequencing was performed using *Sbf*I following a
690 standard protocol [84]. RAD library preparation and sequencing was conducted
691 using a 100bp single-end Illumina HiSeq by Floragenex (Oregon, USA).
692 Accession numbers, sample names and locations for all genome and
693 RAD-seq samples are listed in Table S6.
694

695 **Whole genome alignment and variant calling**. Sequence reads were mapped
696 to the stickleback reference genome using CLC Genome 8.0 as described
697 previously [40]. Alignments were exported as bam files and were sorted and
698 indexed using samtools 1.2 [85]. Variant calling was carried out in two different
699 phases. The first phase was used to call consensus bases at all sites (i.e. variant
700 and invariant) across the genome for all 27 resequenced individuals and the
701 outgroup (*G. wheatlandi*). Mapped reads from all individuals were piled-up using
702 samtools *mpileup* and called against the stickleback BROAD reference genome
703 using the bcftools 1.2 consensus caller without any filters. This consensus call
704 produced a vcf file with a base call for every position in the genome for all
705 samples (27 + 1 outgroup). Consensus calls from this phase were used in later
706 demographic inference using PSMC and ABC. The aim of the second, more
707 stringent variant calling phase was to produce a subset of high-quality
708 polymorphic SNPs with which to examine genome-wide differentiation between
709 the Japan Sea, Pacific and Atlantic Ocean lineages. We used bcftools to filter the
710 consensus-call vcf for these three lineages, only retaining sites with a Phred
711 Quality score >10, and with a maximum individual read depth of 200. Additional
712 downstream filters were applied prior to estimating population genomic
713 parameters from this subset vcf (see below).
714
715 **Mitochondrial genome divergence**. To estimate divergence times based on
716 mitochondrial DNA, we performed Bayesian coalescent analysis using BEAST
717 v2.2.1 [86]. From our resequencing data, we extracted the whole mitochondrial
718 genome from the 26 Japan Sea, Pacific and Atlantic Ocean individuals. We also
719 downloaded two *G. wheatlandi* whole mitogenomes as outgroups (NCBI

720 accession numbers: AB445129 & NC011570). Note that due to poor sequence
721 coverage across the mitogenome we excluded our own re-sequenced *G.*
722 *wheatlandi* individual here. Mitogenomes were aligned using MUSCLE v3.81.3
723 [87] resulting in a 16,549 bp final alignment.

724 Although there is a considerable three-spined stickleback fossil record,
725 it is unfortunately of little use for providing fossil calibration dates for splits
726 within the *Gasterosteus* genus [88,89]. However biogeographical events can also
727 be used to calibrate node estimates and as such we used a normal prior (mean =
728 1.5 million years, SD = 0.75 million years) on the split between the Japan Sea and
729 Pacific Ocean *G.aculeatus* lineages. We provided a further normal prior on the
730 split date between the Pacific and Atlantic Ocean mitochondrial lineages (mean =
731 0.5 million years, SD = 0.25 million years). The latter prior distribution was
732 intentionally made wide to reflect uncertainty surrounding this estimate. Initial
733 analyses with BEAST indicated that marginal prior distributions for node ages
734 did not behave as specified in the model and instead returned extremely recent
735 divergence times with low likelihood support. This is a common bias in
736 coalescent divergence time dating and use of a calibrated prior removed this
737 issue [90,91]. As a result, we performed all further analyses with a calibrated
738 Yule prior. Incorrect choice of molecular clock model can seriously bias
739 coalescent estimates of lineage divergence times and so care must be taken to
740 ensure the appropriate model is chosen [92,93]. We used path-sampling analysis
741 in BEAST to estimate model marginal likelihoods for three different clock models
742 – strict, relaxed lognormal and relaxed exponential. For each model, Markov
743 chain Monte Carlo (MCMC) was run for 5 x 10^7 with 60 steps and marginal
744 likelihoods were calculated using BEAST. We then ran the final model using two

745 10^8 independent MCMC runs. Runs were assessed in TRACER [94] to ensure
746 convergence and that ESS values > 200 – i.e. the posterior was adequately
747 sampled. Independent runs were then combined to produce posterior estimates
748 of divergence times and substitution rates.
749
750 **Nuclear phylogenetic analysis and genealogical sorting index (*gsi*).** To
751 investigate nuclear phylogenetic discordance, we constructed maximum
752 likelihood trees from consensus sequences for non-overlapping 10, 50 and 100
753 kb sliding windows following Martin et al [28]. The best-fit tree was estimated
754 for each window using RAxML with a 'GTRGAMMA' model and a random number
755 seed [95]. Trees were classified using a custom R script
756 (https://github.com/markravinet) that binned trees based on whether they
757 matched three different topologies; species, geography, ancestral – or were
758 unresolved. For the species category, all Atlantic, Pacific and Japan Sea
759 individuals form separate monophyletic groups; for the geography category,
760 Japan Sea and Pacific Ocean form a monophyletic group separate to the Atlantic
761 Ocean; trees where the Atlantic Ocean grouped monophyletically with the Japan
762 Sea were classed as ancestral. Trees that did not fit any of these categories were
763 classified as unresolved. Following categorisation, trees were then standardised
764 to ensure equal branch lengths using the *compute.brlen* function from the
765 *Phytools* R package [96] and were finally visualised for each gene tree class using
766 the *densiTree* function in the R package *Phangorn* [97].
767 We additionally used the non-overlapping Maximum Likelihood
768 phylogenies to calculate genealogical sorting index (*gsi*) [47]. We used a custom

769 R script to estimate *gsi* across the autosome of 26 resequenced individuals. This
770 allowed us to compare autosomal signals of introgression with a reduction in *gsi*.
771
772 **Population size change over time**. We used PSMC to estimate fluctuations in
773 effective population size over time [98]. PSMC uses the density of heterozygote
774 sites across a single diploid genome to estimate blocks of constant TMRCA that
775 are split by recombination and then uses these to infer ancestral effective
776 population sizes (N_e) over time [61,98]. Since PSMC can only analyse a single
777 genome at a time, we ran the program separately on each of the 26 resequenced
778 genomes from Japan Sea, Pacific and Atlantic Ocean lineages. We additionally ran
779 the analyses for a reseqeuenced genome of a marine ecotype fish from Little
780 Campbell River, Canada as representatives of the Eastern Pacific. Consensus
781 sequences for each genome were converted to PSMC format - a binary format
782 indicating the presence/absence of heterozygous sites within a specified
783 window. We used 100 bp windows, requiring a minimum of 10,000 'good' sites
784 to be present on a genome scaffold in order for it to be included; heterozygous
785 sites with a Phred Quality score <20 were removed. We then ran PSMC for 30
786 iterations with a maximum coalescent time of 15 (measured in units of $2N_0$
787 where N_0 is ancestral population size). Due to the difficulty of inferring past
788 effective population sizes across this time, PSMC requires the user to provide
789 intervals which are combined to produce the same effective population size [98].
790 Since this method is least accurate for recent (i.e. < 20 kyr BP) and more ancient
791 periods [98], we estimated N_e for 47 intervals, combining the first four and the
792 last three using the command "4+19*2+4". To scale our results from coalescent
793 units, we assumed a generation time of 1 year [99] and used an autosomal

794 substitution rate of 7.1 x 10^{-9} [100]. Finally, to provide confidence intervals for
795 our N_e estimates, we performed 100 bootstraps on 500 kb segments for each
796 analysis.
797
798 **Approximate Bayesian Computation (ABC)**. We used ABC to test different
799 scenarios of divergence between the Japan Sea and Pacific Ocean lineage and to
800 estimate demographic parameters, such as divergence time and migration rate,
801 under these scenarios.
802 To obtain loci suitable for our ABC analysis, we randomly sampled
803 nuclear loci from the 20 resequenced genomes (sympatric Japan Sea and Pacific
804 Ocean) using a similar approach to Nadachowska-Brzyska et al [14]. Using a
805 custom R script, we produced a bed file of reference genome coordinates for 2 kb
806 loci randomly sampled at 125 kb intervals; resulting in 2,378 loci potential per
807 individual. Using a custom python script, we called consensus sequences for each
808 locus from the consensus vcf. This script created two haplotype sequences for
809 each of the 2 kb loci, randomly assigning heterozygous variants to one of the two
810 called haplotypes; this step allowed us to use unphased data for demographic
811 analyses [66,101]. We then further filtered these loci to include only those that
812 occur on autosomes, with >1,000 bp sequence and a minimum of 30% coverage
813 (i.e. > 70% bases were called) for all 20 individuals. This resulted in a final
814 dataset of 1,874 loci. Functions and scripts for generating coordinates and
815 extracting and filtering consensus sequences are available on GitHub
816 (https://github.com/markravinet/genome_sampler).
817 Following Robinson et al [66] we used a custom R-based control script
818 and msABC [102] to perform simulations, calculate summary statistics and

819 quantify their distribution across the genome in a single step. This approach
820 offers considerable flexibility in establishing prior probability distributions for
821 each of the estimated parameters. Furthermore, given the large size of our
822 dataset (i.e. approximately 2,000 loci for 20 individuals) each simulation
823 produces a large amount of data, making storage a challenge. Using R to interface
824 with msABC allowed us to greatly reduce the required data storage.

825 For each of the 15 models we performed 10^6 simulations. We used a
826 combination of GNU Parallel [103] and independent runs across multiple
827 computing cores to reduce analysis speed to approximately 1 day per model
828 (scripts and additional instructions available on Github:
829 https://github.com/markravinet).

830 We initially ran our simulations to produce all the available summary
831 statistics that msABC calculates. However since summary statistic choice can
832 greatly alter the outcomes of ABC analyses [104,105], all post-simulation ABC
833 analyses were conducted using subsets of 29, 20 and 12 summary statistics.
834 Following completion of the simulation step, we performed a neural-network
835 rejection step on log-transformed parameter estimates with a tolerance of 0.01
836 using the abc function in the R package *abc* [106]. Posterior probability was
837 estimated for each model using the R *abc postpr* function with multinomial
838 logistic regression for a range of tolerance values representing 0.1%, 0.5%, 1%
839 and 3% of the simulated data (i.e. 1,000, 5,000, 10,000 and 30,000 datasets
840 respectively). In keeping with a hierarchical analysis [14], we performed two
841 rounds of model selection. We first chose the growth model with the highest
842 posterior probability within each divergence scenario. Following this, we

843 performed model selection on the five models with the highest support within
844 each divergence category.
845 In order to ensure our ABC approach was reliable, we used pseudo-
846 observed datasets (PODs) to assess how well we could discriminate between
847 different divergence scenarios. Essentially, this involves randomly selecting a
848 series of simulated dataset from a known model (hence pseudo-observed) and
849 then rerunning the model selection procedure to see whether the true model
850 could be recovered. For further details of our POD-based sensitivity analysis and
851 ABC approach, see the *Supplementary methods* section.
852
853 **Detecting genome-wide divergence and recent introgression**. Weir and
854 Cockerham's F_{ST} [107] was calculated using 10 and 50 kb non-overlapping
855 windows with VCFtools 0.113 [108]. To calculate haplotype-based statistics such
856 as d_{XY}, G_{MIN} and f_d, we used a modified version of a python script used by Martin
857 et al [28]. In addition to our main filters on the dataset (see *Genome Alignment*
858 *and Variant calling*), we only calculated these haplotype-based statistics for
859 windows with >50% of useable bases – i.e. >5,000 sites within a 10 kb sliding
860 window. For autosomal statistics, all individuals were included in the analyses.
861 For comparing the ancestral (chrXIX) and neo-sex chromosomes (chrIX) with
862 autosomes (Fig. 5E and 5F), we re-ran the analyses of all chromosomes using
863 only females. In addition to 10 kb windows, we also performed analyses for non-
864 overlapping 50 kb windows to aid visualisation; the results from all analyses
865 were then combined into a single dataset using custom R scripts.
866 We calculated recently established statistics, G_{MIN} and f_d, for detecting
867 introgression between divergent lineages [26,54]. G_{MIN} is particularly suited for

868 identifying recent, low frequency introgression [26] whereas f_d can also identify
869 stronger, high frequency introgression events [54]. Importantly, both methods
870 are robust to variation in recombination rate variation. Initial genome scans
871 conducted using G_{MIN} revealed a series of 'valleys' present across the genome.
872 Detection of such valleys, like genomic islands of divergence, presents a variety
873 of methodological issues. Firstly, how do we determine that G_{MIN} valleys are not
874 due to stochastic variation in genealogy amongst loci? Secondly, how do we
875 measure the size and distribution of valleys of introgression? Finally, how can we
876 determine a null or expected distribution of valleys across the genome to test for
877 the under- or overrepresentation of valleys?

878 To deal with each of these issues in turn, we first performed
879 chromosome-specific permutations to identify the null distribution of the value
880 of G_{MIN}. Specifically, we shuffled the nucleotide sequence of each chromosome
881 100 times and estimated G_{MIN} for 11 different sliding window sizes (5,000, 5,500,
882 6,000, 6,500, 7,000, 7,500, 8,000, 8,500, 9,000, 9,500 and 10,000 kb),
883 representing the distribution of useable sites from the empirical dataset. We
884 then used the lower 99 percentile of the permutations to determine the mean
885 value of G_{MIN} below which a window could be classified as a valley. Identifying
886 the boundaries of divergent genome regions is somewhat subjective and open to
887 potential bias [109]. To account for this, we used a hidden Markov-model (HMM)
888 approach to classify windows into two states - i.e. valleys or non-valleys – and to
889 estimate the probability of state transition. Following Soria-Caracasco *et al.* [55],
890 we used the R package HiddenMarkov [110] on a logit transformed G_{MIN}
891 distribution. Transition probabilities between the two states were symmetrical
892 with an emphasis on it being difficult to transition between states (p = 0.1) but

893 relatively easy to remain within a state (p = 0.9). Since valleys are relatively rare
894 in the genome, we set our models to start in the non-valley state and we
895 provided estimated parameter values for the states based on the empirical
896 distribution. Our permutation test to find the 99 percentile therefore also
897 provided an empirical basis for identifying valleys. HMM estimates were run for
898 both the sympatric and allopatric comparisons using the *baumwelch* function to
899 estimate parameters using the Baum-Welch algorithm and the *viterbi* function to
900 estimate the sequence of states using the Viterbi algorithm. We used a similar
901 approach to identify f_d peaks but we instead performed the analysis using
902 untransformed f_d values only in the sympatric Japanese *G. aculeatus* and *G.*
903 *nipponicus* comparison.
904
905 **RAD-seq data processing and admixture analysis**. RAD sequence reads were
906 demultiplexed and processed using the *process_radtags* module of Stacks 1.30
907 [111]. All reads were trimmed to 90 bp and any read where the average Phred
908 quality score dropped below 10 in a 9 bp sliding window was discarded.
909 Following filtering, reads were mapped to the Roesti *et al.* [59] build of the *G.*
910 *aculeatus* genome using GSNAP [112] allowing a maximum of two indels to be
911 present in an alignment, reporting no suboptimal hits, allowing a maximum of 8
912 mismatches and printing only the best alignment. SNPs were then called using
913 the samtools and bcftools pipeline [113]. Called variants were then filtered using
914 vcftools to remove all sites with greater than 25% missing data, to include
915 genotypes only with an individual depth between 15X and 100X, to remove all
916 sites with a Phred quality score below 20 and with a minor allele frequency
917 below 0.05. Since common admixture analyses assume independence among

918 sites (i.e. the absence of linkage disequilibrium) [114], we additionally pruned
919 our RAD-derived SNP dataset using plink [115] , removing all sites where
920 pairwise linkage disequilibrium was greater than 0.4.
921 PCA on allele frequencies from all individuals was conducted using the
922 *glPca* function from the R package adegenet [116]. Admixture analysis was
923 carried out on a subset of samples from the Japanese archipelago using
924 STRUCTURE [50,51]. For each value of *K* from 1 to 8, the program was run for 10
925 iterations with a burn-in of 10,000 steps followed by 20,000 MCMC steps. The
926 most likely value of *K* was assessed using STRUCTURE HARVESTER [117].
927
928 **Detecting the direction of introgression**. We investigated the direction of gene
929 flow between the Japan Sea and Pacific Ocean lineages using partitioned *D*
930 statistics [56]. This is conceptually similar to standard four population ABBA-
931 BABA tests for gene flow but includes a fifth population – an allopatric lineage of
932 the Japan Sea. This balances the assumed phylogeny (i.e. ((P1, P2), ($P3_1$, $P3_2$), O)
933 and therefore allows us to rotate the populations used in the analysis – i.e.
934 testing for an enrichment of gene flow in both directions. We therefore tested
935 two topologies ((AT, PO), (JS_S, JS_A), O) and ((JS_A, JS_S), (PO, AT), O) (see Fig S14).
936 For either test topology, an excess of the ABBAA (compared to BABAA) or
937 ABBBA (compared to BABBA) positions in a genome window inflates partitioned
938 D statistics above zero – indicating gene flow from the P3 into P2.
939 Given that the partitioned *D* approach has attracted some criticism, we
940 also calculated D_{FOIL} statistics [57]. D_{FOIL} is an additional extension of the four
941 population test but one that incorporates all possible introgression events for a

942 symmetric four population tree (excluding the outgroup). We used the same test
943 phylogeny as with the partitioned D statistics (see Figs S14 & S15).
944 Both partitioned D and D_{FOIL} are based on ABBA/BABA methods – i.e.
945 where only a single individual is present at the tips of the phylogeny. To account
946 for this, we extended both methods to account for allele frequency data, meaning
947 our site pattern counts are weighted by allele frequencies [54]. To calculate both
948 D and D_{FOIL} statistics, we used a modified version of a python script used by
949 Martin et al [28].
950
951 **Characterization of introgression sites**. In order to characterize regions of
952 introgression, we identified candidate regions showing a strong signature of
953 introgression (i.e. G_{MIN} valleys and f_{d} peaks) from our genome scan approach. We
954 then counted the number of unique genes falling within our candidate
955 valleys/peaks and compared this to a null distribution generated by 1,000
956 random samples of 10 kb non-valley/non-peak genome windows for the same
957 number and size range as the valleys or peaks.
958 We then tested whether genes in introgressed regions were more likely
959 to have any specific functions. To achieve this, we used gene ontology (GO)
960 analysis on genes in valleys and 1,000 randomly chosen from across the genome.
961 GO analysis was performed with the ClueGO plugin [118] for Cytoscape 3.4.0
962 [119]. Since functional annotations for this analysis were drawn from the human
963 genome, we first generated a list of human-stickleback orthologous gene IDs
964 (Ensembl Biomart 86). We then subset our candidate and random gene sets to
965 include only orthologous genes. Several human genes have multiple stickleback
966 orthologs; we therefore allowed only a single, randomly chosen occurrence of

967 each human gene in both sets to prevent pseudo-replication. A hypergeometric
968 test was conducted for testing enrichment with Benjamini & Hochberg FDR
969 correction [120].
970

971 ## Acknowledgements

972 We are grateful to Manabu Kume, Seiichi Mori, and staff at the Aquarium Gobius
973 for providing samples and Katsushi Yamaguchi for technical assistance. Keisuke
974 Honda, the Institute of Statistical Mathematics and the DNA Data Bank of Japan
975 are thanked for their help with running analyses on supercomputers. We are
976 additionally grateful to Simon Martin, John Robinson and David Marques for
977 sharing scripts and their advice on analyses. Freddy Chain and Philine Feulner
978 are also thanked for their assistance with their sequence data. We thank
979 Cassandra Trier for her assistance and advice with GO analyses. All members of
980 the Kitano Lab provided invaluable advice throughout the project. We would also
981 like to thank Mark Kirkpatrick and members of his lab for comments on
982 manuscript.
983

984 ## References

985 1. Mayr E. Animal species and evolution. Cambridge, Massachusetts: Havard
986 University Press; 1963.
987 2. Coyne JA, Orr HA. Speciation. New York: Sinaeur; 2004.
988 3. Nosil P. Ecological Speciation. Oxford, UK: Oxford University Press; 2012.
989 4. Wu C-I. The genic view of the process of speciation. J Evol Biol. 2001;14:
990 851–865.

991 5. Feder JL, Egan SP, Nosil P. The genomics of speciation-with-gene-flow.
992 Trends Genet. Elsevier Ltd; 2012;28: 342–350.
993 doi:10.1016/j.tig.2012.03.009
994 6. Nosil P, Feder JL. Genomic divergence during speciation: causes and
995 consequences. Philos Trans R Soc London Ser B. 2012;367: 332–342.
996 doi:10.1098/rstb.2011.0263
997 7. Ravinet M, Faria R, Butlin RK, Galindo J, Bierne N, Rafajlović M, et al.
998 Interpreting the genomic landscape of speciation: finding barriers to gene
999 flow. J Evol Biol. 2017;in press.
1000 8. Nosil P, Feder JL, Flaxman SM, Gompert Z. Tipping points in the dynamics
1001 of speciation. Nat Ecol Evol. Macmillan Publishers Limited; 2017;1: 1–8.
1002 doi:10.1038/s41559-016-0001
1003 9. Riesch R, Muschick M, Lindtke D, Villoutreix R, Comeault AA, Farkas TE, et
1004 al. Transitions between phases of genomic differentiation during stick-
1005 insect speciation. Nat Ecol Evol. Macmillan Publishers Limited, part of
1006 Springer Nature.; 2017;1: 82. doi:10.1038/s41559-017-0082
1007 10. Feder JL, Nosil P, Wacholder AC, Egan SP, Berlocher SH, Flaxman SM.
1008 Genome-wide congealing and rapid transitions across the speciation
1009 continuum during speciation with gene flow. J Hered. 2014;105: 810–820.
1010 doi:10.1093/jhered/esu038
1011 11. Seehausen O, Butlin RK, Keller I, Wagner CE, Boughman JW, Hohenlohe P a,
1012 et al. Genomics and the origin of species. Nat Rev Genet. Nature Publishing
1013 Group; 2014;15: 176–92. doi:10.1038/nrg3644
1014 12. Wolf JB, Ellegren H. Making sense of genomic islands of differentiation in
1015 light of speciation. Nat Rev Genet. Nature Publishing Group; 2017;18: 87–

1016 100. doi:10.1038/nrg.2016.133
1017 13. Burri R, Nater A, Kawakami T, Mugal CF, Olason PI, Smeds L, et al. Linked
1018 selection and recombination rate variation drive the evolution of the
1019 genomic landscape of differentiation across the speciation continuum of
1020 Ficedula flycatchers. Genome Res. 2015;25: 1656–1665.
1021 doi:10.1101/gr.196485.115
1022 14. Nadachowska-Brzyska K, Burri R, Olason PI, Kawakami T, Smeds L,
1023 Ellegren H. Demographic Divergence History of Pied Flycatcher and
1024 Collared Flycatcher Inferred from Whole-Genome Re-sequencing Data.
1025 Payseur BA, editor. PLoS Genet. 2013;9: e1003942.
1026 doi:10.1371/journal.pgen.1003942
1027 15. Via S. Natural selection in action during speciation. Proc Natl Acad Sci.
1028 2009;106: 9939–9946.
1029 16. Marques DA, Lucek K, Meier JI, Mwaiko S, Wagner CE, Excoffier L, et al.
1030 Genomics of Rapid Incipient Speciation in Sympatric Threespine
1031 Stickleback. PLOS Genet. 2016;12: e1005887.
1032 doi:10.1371/journal.pgen.1005887
1033 17. Lawniczak MKN, Emrich SJ, Holloway a K, Regier a P, Olson M, White B, et
1034 al. Widespread divergence between incipient Anopheles gambiae species
1035 revealed by whole genome sequences. Science. 2010;330: 512–4.
1036 doi:10.1126/science.1195755
1037 18. Andrew RL, Rieseberg LH. Divergence is focused on few genomic regions
1038 early in speciation: incipient speciation of sunflower ecotypes. Press.
1039 2013; doi:10.1111/evo.12106.2013
1040 19. Flaxman SM, Wacholder AC, Feder JL, Nosil P. Theoretical models of the

1041 influence of genomic architecture on the dynamics of speciation. Mol Ecol.
1042 2014;23: 4074–4088. doi:10.1111/mec.12750
1043 20. Bierne N, Gagnaire PA, David P. The geography of introgression in a patchy
1044 environment and the thorn in the side of ecological speciation. Curr Zool.
1045 2013;59: 72–86.
1046 21. Cruickshank TE, Hahn MW. Reanalysis suggests that genomic islands of
1047 speciation are due to reduced diversity, not reduced gene flow. Mol Ecol.
1048 2014;23: 3133–3157. doi:10.1111/mec.12796
1049 22. Noor MAF, Bennett SM. Islands of speciation or mirages in the desert?
1050 Examining the role of restricted recombination in maintaining species.
1051 Heredity (Edinb). 2009;103: 439–444.
1052 23. Nater A, Burri R, Kawakami T, Smeds L, Ellegren H. Resolving evolutionary
1053 relationships in closely related species with whole-genome sequencing
1054 data. Syst Biol. 2015;64: 1000–1017. doi:10.1093/sysbio/syv045
1055 24. Rosenzweig BK, Pease JB, Besansky NJ, Hahn MW. Powerful methods for
1056 detecting introgressed regions from population genomic data. Mol Ecol.
1057 2016;25: 2387–2397. doi:10.1111/mec.13610
1058 25. Geneva A, Garrigan D. Population Genomics of Secondary Contact. Genes
1059 (Basel). 2010;1: 124–142. doi:10.3390/genes1010124
1060 26. Geneva AJ, Muirhead CA, Kingan SB, Garrigan D. A new method to scan
1061 genomes for introgression in a secondary contact model. PLoS One.
1062 2015;10: e0118621. doi:10.1371/journal.pone.0118621
1063 27. Meier JI, Marques DA, Mwaiko S, Wagner CE, Excoffier L, Seehausen O.
1064 Ancient hybridization fuels rapid cichlid fish adaptive radiations. Nat
1065 Commun. Nature Publishing Group; 2017;8: 14363.

1066 doi:10.1038/ncomms14363
1067 28. Martin SH, Dasmahapatra KK, Nadeau NJ, Salazar C, Walters JR, Simpson F,
1068 et al. Genome-wide evidence for speciation with gene flow in Heliconius
1069 butterflies. Genome Res. 2013;23: 1817–28. doi:10.1101/gr.159426.113
1070 29. Martin CH, Crawford JE, Turner BJ, Simons LH. Diabolical survival in Death
1071 Valley: recent pupfish colonization, gene flow and genetic assimilation in
1072 the smallest species range on earth. Proc R Soc B Biol Sci. 2016;283:
1073 20152334. doi:10.1098/rspb.2015.2334
1074 30. Hendry AP, Bolnick DI, Berner D, Peichel CL. Along the speciation
1075 continuum in sticklebacks. J Fish Biol. 2009;75: 2000–2036.
1076 31. McKinnon JS, Rundle HD. Speciation in nature: the threespine stickleback
1077 model systems. Trends Ecol Evol. 2002;17: 480–481. doi:10.1016/S0169-
1078 5347(02)02579-X
1079 32. Feulner PGD, Chain FJJ, Panchal M, Huang Y, Eizaguirre C, Kalbe M, et al.
1080 Genomics of Divergence along a Continuum of Parapatric Population
1081 Differentiation. PLOS Genet. 2015;11: e1004966.
1082 doi:10.1371/journal.pgen.1004966
1083 33. Roesti M, Hendry AP, Salzburger W, Berner D. Genome divergence during
1084 evolutionary diversification as revealed in replicate lake-stream
1085 stickleback population pairs. Mol Ecol. 2012;21: 2852–2862.
1086 doi:10.1111/j.1365-294X.2012.05509.x
1087 34. Roesti M, Kueng B, Moser D, Berner D. The genomics of ecological
1088 vicariance in threespine stickleback fish. Nat Commun. Nature Publishing
1089 Group; 2015;6: 8767. doi:10.1038/ncomms9767
1090 35. Wu C-I, Ting C-T. Genes and speciation. Nat Rev Genet. 2004;5: 114–22.

1091 doi:10.1038/nrg1269
1092 36. Kitano J, Ross JA, Mori S, Kume M, Jones FC, Chan YF, et al. A role for neo-
1093 sex chromosomes in stickleback speciation. Nature. 2009;461: 1079–1083.
1094 doi:10.1002/ece3.234
1095 37. Kitano J, Mori S, Peichel CL. Phenotypic divergence and reproductive
1096 isolation between sympatric forms of Japanese threespine sticklebacks.
1097 Biol J Linn Soc. 2007;91: 671–685. doi:10.1111/j.1095-8312.2007.00824.x
1098 38. Higuchi M, Sakai H, Goto A. A new threespine stickleback, Gasterosteus
1099 nipponicus sp. nov. (Teleostei: Gasterosteidae), from the Japan Sea region.
1100 Ichthyol Res. 2014; 1–2. doi:10.1007/s10228-014-0403-1
1101 39. Higuchi M, Goto A. Genetic evidence supporting the existence of two
1102 distinct species in the genus *Gasterosteus* around Japan. Environ Biol
1103 Fishes. 1996;47: 1–16. doi:10.1007/BF00002375
1104 40. Yoshida K, Makino T, Yamaguchi K, Shigenobu S, Hasebe M, Kawata M, et
1105 al. Sex Chromosome Turnover Contributes to Genomic Divergence
1106 between Incipient Stickleback Species. Zhang J, editor. PLoS Genet.
1107 2014;10: e1004223. doi:10.1371/journal.pgen.1004223
1108 41. Kume M, Kitano J, Mori S, Shibuya T. Ecological divergence and habitat
1109 isolation between two migratory forms of Japanese threespine stickleback
1110 (Gasterosteus aculeatus). J Evol Biol. 2010;23: 1436–1446.
1111 doi:10.1111/j.1420-9101.2010.02009.x
1112 42. Ravinet M, Takeuchi N, Kume M, Mori S, Kitano J. Comparative Analysis of
1113 Japanese Three-Spined Stickleback Clades Reveals the Pacific Ocean
1114 Lineage Has Adapted to Freshwater Environments while the Japan Sea Has
1115 Not. Craft JA, editor. PLoS One. 2014;9: e112404.

1116 doi:10.1371/journal.pone.0112404
1117 43. Lackey AC, Boughman JW. Evolution of reproductive isolation in
1118 stickleback fish. Evolution (N Y). 2017;71: 357–371.
1119 doi:10.1038/hdy.2008.69
1120 44. Ortí G, Bell MA, Reimchen TE, Meyer A. Global survey of mitochondrial
1121 DNA sequences in the threespine stickleback: evidence for recent
1122 migrations. Evolution (N Y). 1994;48: 608–622.
1123 45. Yamada M, Higuchi M, Goto A. Extensive introgression of mitochondrial
1124 DNA found between two genetically divergent forms of threespine
1125 stickleback, *Gasterosteus aculeatus*, around Japan. Environ Biol Fishes.
1126 2001;61: 269–284.
1127 46. Feulner PGD, Chain FJJ, Panchal M, Eizaguirre C, Kalbe M, Lenz TL, et al.
1128 Genome-wide patterns of standing genetic variation in a marine
1129 population of three-spined sticklebacks. Mol Ecol. 2012; no-no.
1130 doi:10.1111/j.1365-294X.2012.05680.x
1131 47. Cummings MP, Neel MC, Shaw KL. A genealogical approach to quantifying
1132 lineage divergence. Evolution. 2008;62: 2411–22. doi:10.1111/j.1558-
1133 5646.2008.00442.x
1134 48. Toews DPL, Brelsford A. The biogeography of mitochondrial and nuclear
1135 discordance in animals. Mol Ecol. 2012;21: 3907–30. doi:10.1111/j.1365-
1136 294X.2012.05664.x
1137 49. Catchen J, Bassham S, Wilson T, Currey M, O'Brien C, Yeates Q, et al. The
1138 population structure and recent colonization history of Oregon threespine
1139 stickleback determined using restriction-site associated DNA-sequencing.
1140 Mol Ecol. 2013;22: 2864–2883. doi:10.1111/mec.12330

1141 50. Falush D, Stephens M, Pritchard JK. Inference of population structure:
1142 Extensions to linked loci and correlated allele frequencies. Genetics.
1143 2003;164: 1567–1587.
1144 51. Pritchard JK, Stephens M, Rosenberg NA, Donnelly P. Association mapping
1145 in structured populations. Am J Hum Genet. 2000;67.
1146 52. Kusakabe M, Ishikawa A, Ravinet M, Yoshida K, Makino T, Toyoda A, et al.
1147 Genetic basis for variation in salinity tolerance between stickleback
1148 ecotypes. Mol Ecol. 2016; doi:10.1111/mec.13875
1149 53. Charlesworth B, Morgan MT, Charlesworth D. The effect of deleterious
1150 mutations on neutral molecular variation. Genetics. 1993;134: 1289–303.
1151 Available:
1152 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1205596&to
1153 ol=pmcentrez&rendertype=abstract
1154 54. Martin SH, Davey JW, Jiggins CD. Evaluating the use of ABBA-BABA
1155 statistics to locate introgressed loci. Mol Biol Evol. 2014;32: 244–257.
1156 doi:10.1101/001347
1157 55. Soria-Carrasco V, Gompert Z, Comeault AA, Farkas TE, Parchman TL,
1158 Johnston JS, et al. Stick Insect Genomes Reveal Natural Selection's Role in
1159 Parallel Speciation. Science (80-). 2014;344: 738–742.
1160 doi:10.1126/science.1252136
1161 56. Eaton DAR, Ree RH. Inferring phylogeny and introgression using RADseq
1162 data: An example from flowering plants (Pedicularis: Orobanchaceae). Syst
1163 Biol. 2013;62: 689–706. doi:10.1093/sysbio/syt032
1164 57. Pease JB, Hahn MW. Detection and Polarization of Introgression in a Five-
1165 Taxon Phylogeny. Syst Biol. 2015;64: 651–662.

1166 doi:10.1093/sysbio/syv023
1167 58. Cassidy L, Ravinet M, Mori S, Kitano J. Are Japanese freshwater populations
1168 of threespine stickleback derived from the Pacific Ocean lineage? Evol Ecol
1169 Res. 2013;15: 295–311.
1170 59. Roesti M, Moser D, Berner D. Recombination in the threespine stickleback
1171 genome - Patterns and consequences. Mol Ecol. 2013;22: 3014–3027.
1172 doi:10.1111/mec.12322
1173 60. Li H, Durbin R. Inference of human population history from individual
1174 whole-genome sequences. Nature. Nature Publishing Group; 2011;475:
1175 493–6. doi:10.1038/nature10231
1176 61. Schiffels S, Durbin R. Inferring human population size and separation
1177 history from multiple genome sequences. Nat Genet. Nature Publishing
1178 Group; 2014;46: 919–925. doi:10.1038/ng.3015
1179 62. Kume M, Kitamura T, Takahashi H, Goto A. Distinct spawning migration
1180 patterns in sympatric Japan Sea and Pacific Ocean forms of threespine
1181 stickleback *Gasterosteus aculeatus*. Icthyological Res. 2005;52: 189–193.
1182 doi:10.1007/s10228-005-0269-3
1183 63. Kume M, Kuwahara T, Arai T, Okamoto M, Goto A. A part of the Japan Sea
1184 form of the threespine stickleback, Gasterosteus aculeatus, spawns in the
1185 seawater tidal pools of western Hokkaido Island, Japan. Environ Biol
1186 Fishes. 2006;77: 169–175. doi:10.1007/s10641-006-9068-6
1187 64. Wirtz P. Mother species–father species: unidirectional hybridization in
1188 animals with female choice. Anim Behav. 1999;58: 1–12.
1189 doi:10.1006/anbe.1999.1144
1190 65. Knowles LL. Statistical Phylogeography. Annu Rev Ecol Syst. 2009;40:

1191 593–612.
1192 66. Robinson JD, Bunnefeld L, Hearn J, Stone GN, Hickerson MJ. ABC inference
1193 of multi-population divergence with admixture from un-phased
1194 population genomic data. Mol Ecol. 2014;23: 4458–4471.
1195 doi:10.1111/mec.12881
1196 67. Sousa V, Hey J. Understanding the origin of species with genome-scale
1197 data: modelling gene flow. Nat Rev Genet. Nature Publishing Group;
1198 2013;14: 404–14. doi:10.1038/nrg3446
1199 68. Yamada M, Higuchi M, Goto A. Long-term occurrence of hybrids between
1200 Japan Sea and Pacific Ocean forms of threespine stickleback, *Gasterosteus*
1201 *aculeatus*, in Hokkaido Island, Japan. Environ Biol Fishes. 2007;80: 435–
1202 443.
1203 69. Mäkinen HS, Merilä J. Mitochondrial DNA phylogeography of the three-
1204 spined stickleback (*Gasterosteus aculeatus*) in Europe - Evidence for
1205 multiple glacial refugia. Mol Phylogenet Evol. 2008;46: 167–182.
1206 doi:10.1016/j.ympev.2007.06.011
1207 70. Ravinet M, Harrod C, Eizaguirre C, Prodöhl P a. Unique mitochondrial DNA
1208 lineages in Irish stickleback populations: cryptic refugium or rapid
1209 recolonization? Ecol Evol. 2013; n/a-n/a. doi:10.1002/ece3.853
1210 71. Hohenlohe PA, Bassham S, Etter PD, Stiffler N, Johnson EA, Cresko WA.
1211 Population genomics of parallel adaptation in threespine stickleback using
1212 sequenced RAD tags. PloS Genet. 2010;6: e1000862.
1213 72. Lescak EA, Bassham SL, Catchen J, Gelmond O, Sherbick ML, von Hippel FA,
1214 et al. Evolution of stickleback in 50 years on earthquake-uplifted islands.
1215 Proc Natl Acad Sci U S A. 2015;112: E7204–E7212.

1216 doi:10.1073/pnas.1512020112
1217 73. Barton NH, de Cara MAR. The evolution of strong reproductive isolation.
1218 Evolution. 2009;63: 1171–90. doi:10.1111/j.1558-5646.2009.00622.x
1219 74. Barton N, Bengtsson BO. The barrier to genetic exchange between
1220 hybridising populations. Heredity (Edinb). 1986;57: 357–376.
1221 doi:10.1038/hdy.1986.135
1222 75. Muirhead CA, Presgraves DC. Hybrid Incompatibilities, Local Adaptation,
1223 and the Genomic Distribution of Natural Introgression between Species.
1224 Am Nat. 2016;187: 249–261. doi:10.1086/684583
1225 76. Hedrick PW. Adaptive introgression in animals: examples and comparison
1226 to new mutation and standing variation as sources of adaptive variation.
1227 Mol Ecol. 2013; doi:10.1111/mec.12415
1228 77. Racimo F, Sankararaman S, Nielsen R, Huerta-Sánchez E. Evidence for
1229 archaic adaptive introgression in humans. Nat Rev Genet. Nature
1230 Publishing Group; 2015;16: 359–371. doi:10.1038/nrg3936
1231 78. Castric V, Bechsgaard J, Schierup MH, Vekemans X. Repeated adaptive
1232 introgression at a gene under multiallelic balancing selection. PLoS Genet.
1233 2008;4. doi:10.1371/journal.pgen.1000168
1234 79. Elgvin TO, Trier CN, Tørresen OK, Hagen, Ingerid J, Lien S, Nederbragt AJ,
1235 et al. The genomic mosaicism of hybrid speciation. Sci Adv. 2017;3.
1236 80. Sun C, Huo D, Southard C, Nemesure B, Hennis A, Cristina Leske M, et al. A
1237 signature of balancing selection in the region upstream to the human
1238 UGT2B4 gene and implications for breast cancer risk. Hum Genet.
1239 2011;130: 767–775. doi:10.1007/s00439-011-1025-6
1240 81. Bolnick DI, Stutz WE. Frequency dependence limits divergent evolution by

1241 favouring rare immigrants over residents. Nature. Nature Publishing
1242 Group; 2017;546: 285–288. doi:10.1038/nature22351
1243 82. Ishikawa A, Kusakabe M, Yoshida K, Ravinet M, Makino T, Toyoda A, et al.
1244 Different contributions of local- and distant-regulatory changes to
1245 transcriptome divergence between stickleback ecotypes. 2017; 1–17.
1246 doi:10.1111/evo.13175
1247 83. Ravinet M, Prodöhl PA, Harrod C. On Irish stickleback: morphological
1248 diversification in a secondary contact zone. Evol Ecol Res. 2013;15: 271–
1249 294.
1250 84. Baird N a, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis Z a, et al.
1251 Rapid SNP discovery and genetic mapping using sequenced RAD markers.
1252 PLoS One. 2008;3: e3376. doi:10.1371/journal.pone.0003376
1253 85. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The
1254 Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25:
1255 2078–9. doi:10.1093/bioinformatics/btp352
1256 86. Drummond AJ, Rambaut A. BEAST: Bayesian evolutionary analysis by
1257 sampling trees. BMC Evol Biol. 2007;7: 214–222.
1258 87. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and
1259 high throughput. Nucleic Acids Res. 2004;32: 1792–7.
1260 doi:10.1093/nar/gkh340
1261 88. Bell MA. Paleobiology and evolution of threespine stickleback. In: Bell MA,
1262 Foster SA, editors. The evolutionary biology of the threespine stickleback.
1263 Oxford: Oxford University Press; 1994. pp. 438–471.
1264 89. Bell MA, Stewart JD, Park PJ. The world's oldest fossil threespine
1265 stickleback fish. Copeia. 2009;2009: 256–265.

1266 90. Heled J, Drummond AJ. Calibrated tree priors for relaxed phylogenetics
1267 and divergence time estimation. Syst Biol. 2012;61: 138–49.
1268 doi:10.1093/sysbio/syr087
1269 91. Wheat CW, Wahlberg N. Critiquing blind dating: the dangers of over-
1270 confident date estimates in comparative genomics. Trends Ecol Evol.
1271 Elsevier Ltd; 2013;28: 636–42. doi:10.1016/j.tree.2013.07.007
1272 92. Baele G, Lemey P, Bedford T, Rambaut A, Suchard M a, Alekseyenko A V.
1273 Improving the accuracy of demographic and molecular clock model
1274 comparison while accommodating phylogenetic uncertainty. Mol Biol Evol.
1275 2012;29: 2157–67. doi:10.1093/molbev/mss084
1276 93. Baele G, Li WLS, Drummond AJ, Suchard M a, Lemey P. Accurate model
1277 selection of relaxed molecular clocks in bayesian phylogenetics. Mol Biol
1278 Evol. 2013;30: 239–43. doi:10.1093/molbev/mss243
1279 94. Rambaut A, Drummond AJ. Tracer v1.5. 2009.
1280 95. Stamatakis A. RAxML-VI-HPC: maximum likelihood-based phylogenetic
1281 analyses with thousands of taxa and mixed models. Bioinformatics.
1282 2006;22: 2688–2690. doi:10.1093/bioinformatics/btl446
1283 96. Revell LJ. phytools: an R package for phylogenetic comparative biology
1284 (and other things). Methods Ecol Evol. 2012;3: 217–223.
1285 doi:10.1111/j.2041-210X.2011.00169.x
1286 97. Schliep KP. phangorn: Phylogenetic analysis in R. Bioinformatics. 2011;27:
1287 592–593. doi:10.1093/bioinformatics/btq706
1288 98. Li H, Durbin R. Inference of human population history from individual
1289 whole-genome sequences. Nature. 2011;475: 493–6.
1290 doi:10.1038/nature10231

1291 99. Bell MA, Foster SA. Introduction to the evolutionary biology of the
1292 threespine stickleback. In: Bell MA, Foster SA, editors. The Evolutionary
1293 Biology of the Threespine Stickleback. Oxford: Oxford University Press;
1294 1994. pp. 1–27.
1295 100. Guo B, Chain FJ, Bornberg-Bauer E, Leder EH, Merilä J. Genomic divergence
1296 between nine- and three-spined sticklebacks. BMC Genomics. 2013;14:
1297 756. doi:10.1186/1471-2164-14-756
1298 101. Mailund T, Halager AE, Westergaard M, Dutheil JY, Munch K, Andersen LN,
1299 et al. A new isolation with migration model along complete genomes infers
1300 very different divergence processes among closely related great ape
1301 species. PLoS Genet. 2012;8: e1003125.
1302 doi:10.1371/journal.pgen.1003125
1303 102. Pavlidis P, Laurent S, Stephan W. msABC: a modification of Hudson's ms to
1304 facilitate multi-locus ABC analysis. Mol Ecol Resour. 2010;10: 723–7.
1305 doi:10.1111/j.1755-0998.2010.02832.x
1306 103. Tange O. GNU Parallel - The Command-Line Power Tool. ;login USENIX
1307 Mag. Frederiksberg, Denmark; 2011;36: 42–47. Available:
1308 http://www.gnu.org/s/parallel
1309 104. Csilléry K, Blum MGB, Gaggiotti OE, François O, Csillery K, Francois O.
1310 Approximate Bayesian Computation (ABC) in practice. Trends Ecol Evol.
1311 2010;25: 410–418. doi:10.1016/j.tree.2010.04.001
1312 105. Bertorelle G, Benazzo A, Mona S. ABC as a flexible framework to estimate
1313 demography over space and time: some cons, many pros. Mol Ecol.
1314 2010;19: 2609–2625. doi:10.1111/j.1365-294X.2010.04690.x
1315 106. Csilléry K, François O, Blum MGB. abc: an R package for approximate

1316 Bayesian computation (ABC). Methods Ecol Evol. 2012;3: no-no.
1317 doi:10.1111/j.2041-210X.2011.00179.x
1318 107. Weir B, Cockerham C. Estimating F-Statistics for the Analysis of Population
1319 Structure. Evolution (N Y). 1984;38: 1358–1370. Available:
1320 http://www.jstor.org/stable/2408641
1321 108. Danecek P, Auton A, Abecasis G, Albers C a, Banks E, DePristo M a, et al.
1322 The variant call format and VCFtools. Bioinformatics. 2011;27: 2156–8.
1323 doi:10.1093/bioinformatics/btr330
1324 109. Smadja CM, Canbäck B, Vitalis R, Gautier M, Ferrari J, Zhou J-J, et al. Large-
1325 scale candidate gene scan reveals the role of chemoreceptor genes in host
1326 plant specialization and speciation in the pea aphid. Evolution. 2012;66:
1327 2723–38. doi:10.1111/j.1558-5646.2012.01612.x
1328 110. Harte D. HiddenMarkov: Hidden Markov Models. R package version 1.8-7.
1329 Wellington: Statistics Research Associates; 2016.
1330 111. Catchen J, Hohenlohe P a., Bassham S, Amores A, Cresko W a. Stacks: an
1331 analysis tool set for population genomics. Mol Ecol. 2013;22: 3124–3140.
1332 doi:10.1111/mec.12354
1333 112. Wu TD, Nacu S. Fast and SNP-tolerant detection of complex variants and
1334 splicing in short reads. Bioinformatics. 2010;26: 873–81.
1335 doi:10.1093/bioinformatics/btq057
1336 113. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The
1337 Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25:
1338 2078–9. doi:10.1093/bioinformatics/btp352
1339 114. Pritchard JK, Stephens M, Donnelly P. Inference of population structure
1340 using multilocus genotype data. Genetics. 2000;155: 945–959.

115. 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. doi:10.1086/519795

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

117. 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. doi:10.1007/s12686-011-9548-7

118. 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. doi:10.1093/bioinformatics/btp101

119. 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. doi:10.1101/gr.1239303.metabolite

120. 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

Figures

1366

1367 **Fig 1. The Japan Sea stickleback is a separate species.** (A) Rooted nuclear
1368 consensus tree for Japan Sea, Pacific Ocean and Atlantic Ocean stickleback
1369 lineages from 10 kb non-overlapping sliding windows across the autosome. Red
1370 trees indicate species clustering; blue trees indicate geographical clustering and
1371 green trees reflect ancestral polymorphism. NB: Only 1,000 subsampled species
1372 trees are shown here to aid illustration. (B) Mitogenome Bayesian consensus
1373 tree shows divergence between two mitochondrial clades – all Japanese
1374 sticklebacks (*G. nipponicus* and *G. aculeatus*) and *G. aculeatus* occurring in Europe
1375 and North America. (C) Present day distribution of *G. aculeatus* (blue) and *G.*
1376 *nipponicus* (red) around the Japanese archipelago. The two species overlap in
1377 Hokkaido, Northern Japan and samples for this study were collected in
1378 Bekanbeushi River in Akkeshi unless noted. (D) PSMC plot of 26 resequenced
1379 genomes shows a steady population size in the Pacific Ocean lineage (blue) but a
1380 bottleneck around 0.15-0.3 million years before present and a subsequent
1381 increase in the Japan Sea lineage (orange).

1382

1383 **Fig 2. ABC analysis supports isolation with gene flow.** (A) A model of isolation
1384 with migration and a bottleneck in the Japan Sea lineage is best supported by
1385 ABC analysis using ~2,000 nuclear loci (see Table 1). Posterior probability
1386 densities for model parameters estimated using neural network analysis, a 1%
1387 tolerance rate and 20 summary statistics. Parameters are: T = time of split, m_{12} =
1388 migration from Japan Sea to Pacific Ocean, m_{21} = migration from Pacific Ocean to
1389 Japan Sea; T_G = timing of bottleneck, N_{PO} = Pacific Ocean effective population size,
1390 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) mean migration rates. Figures on each panel are median parameter estimates.

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 [32–34,52].

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.

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) 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.

1415 **Table 1. Posterior probability values for models for final ABC model**
1416 **selection using neural network and standard rejection methods.** All
1417 estimates produced using a tolerance of 1% and 20 summary statistics. Bold text
1418 indicates the model where posterior probability provides the highest support.
1419 Models are I = isolation, IM = isolation with migration, IAM = isolation and
1420 ancient migration, IRM = isolation and recent migration, IARM = isolation with
1421 ancient and recent migration.

Divergence model	Growth model	Standard rejection	Neural network
IM	bottleneck	**0.265**	**0.511**
I	bottleneck	0.130	0.008
IAM	bottleneck	0.151	0.009
IARM	bottleneck	0.230	0.343
IRM	bottleneck	0.224	0.129

1422

1423

1424 **Table 2. Genome-wide averages for measures of divergence and**

1425 **introgression.** F_{ST}, d_{XY} , G_{MIN}, and f_d for all pairwise comparisons of Japan Sea

1426 (JS), Pacific Ocean (PO) and Atlantic Ocean sticklebacks (AT) are shown. Mean ±

1427 SD and lower and upper limits of the 95% confidence interval (in parenthesis)

1428 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

1429

1430 ## Supporting information

1431 **S1 Table. Classification of trees.** Proportion of trees drawn from 10, 50 and
1432 100 kb windows representing species, ancestral and geographical topologies.

1433 **S2 Table. Mitogenome analysis.** Marginal log-likelihood values and Bayes
1434 factor comparisons from path sampling of substitution clock models used for
1435 mitogenome phylogeny and divergence time estimation.

1436 **S3 Table. Median, L95% and U95% HPD (highest probability densities) for**
1437 **demographic parameters estimated under IM + bottleneck model.**

1438 **S4 Table. Number of valleys and valley per Mb for all chromosomes.**

1439 **S5 Table. Enriched GO terms for genes present in G_{MIN} valleys.**

1440 **S6 Table. Enriched GO terms for genes present in f_d peaks.**

1441

1442 **S7 Table. Sample names, locations and accession numbers.**

1443

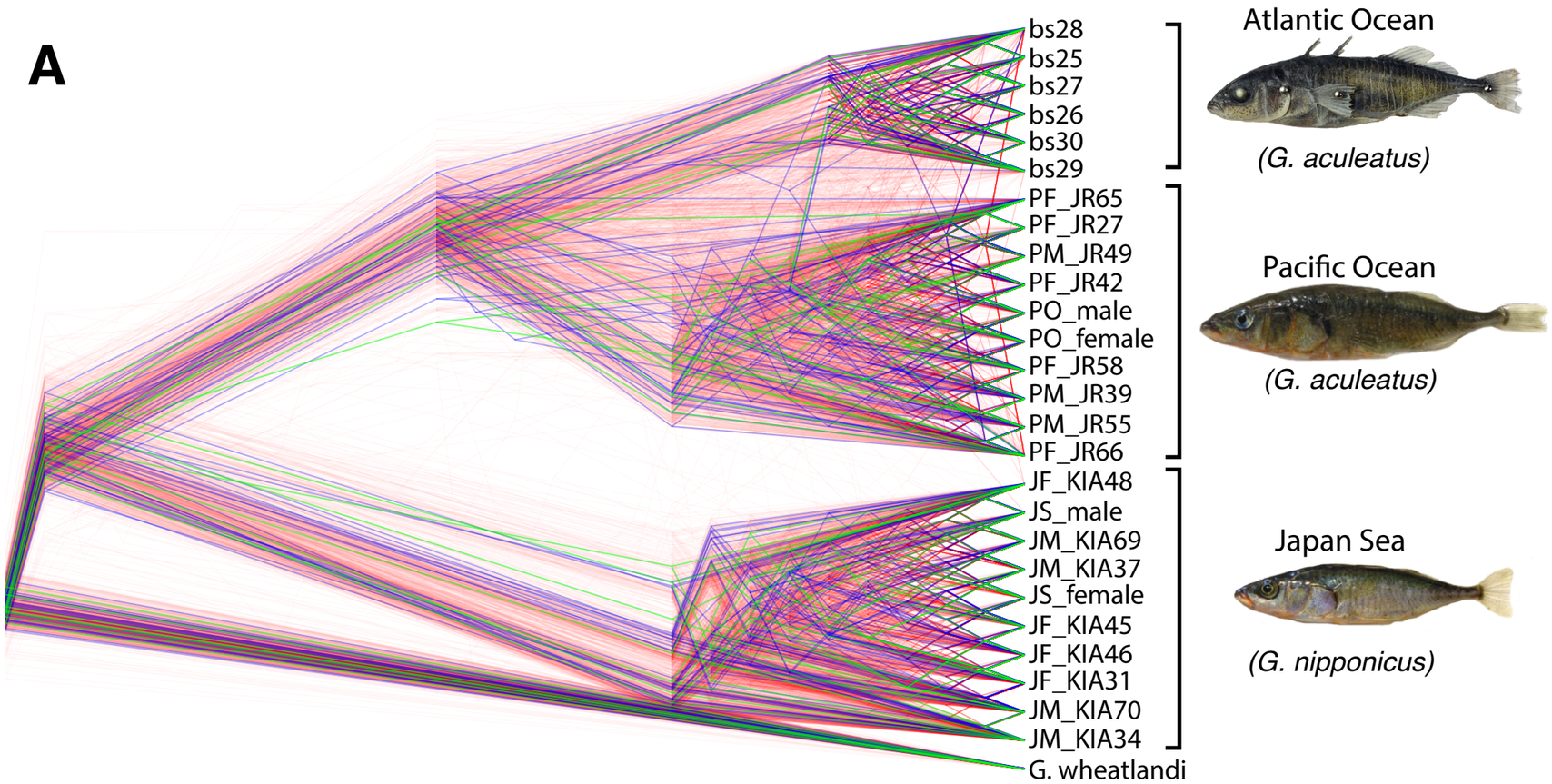
1444 **S1 Fig. Global species distribution.** The global distribution of the Pacific and
1445 Atlantic Ocean lineages of *G. aculeatus* allow sympatric and allopatric
1446 comparisons with *G. nipponicu*; AT = Atlantic Ocean, PO = Pacific Ocean and JS =
1447 Japan Sea.

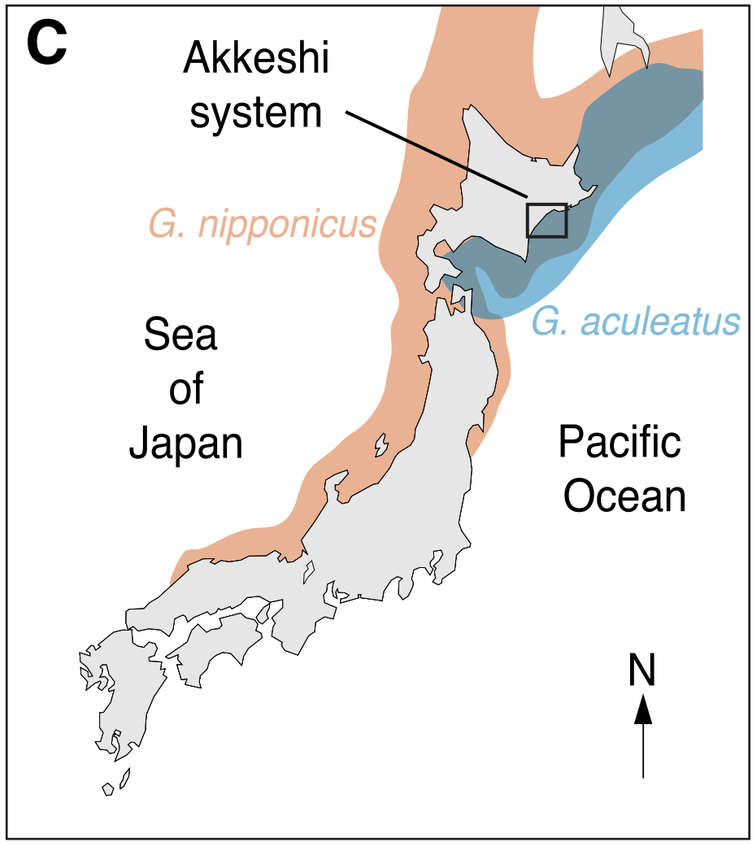
1448 **S2 Fig. Japanese stickleback mitonuclear discordance.** (A) Mitogenome
1449 Bayesian tree shows divergence between two mitochondrial clades – the
1450 Transpacific (TPN) and European North American (ENA); asterisks on nodes
1451 indicate appropriate densities shown in (B). (B) Posterior probability densities
1452 for mitochondrial divergence time between *G. aculeatus* and *G. nipponicus* and
1453 between Pacific and Atlantic populations of the ENA clade.

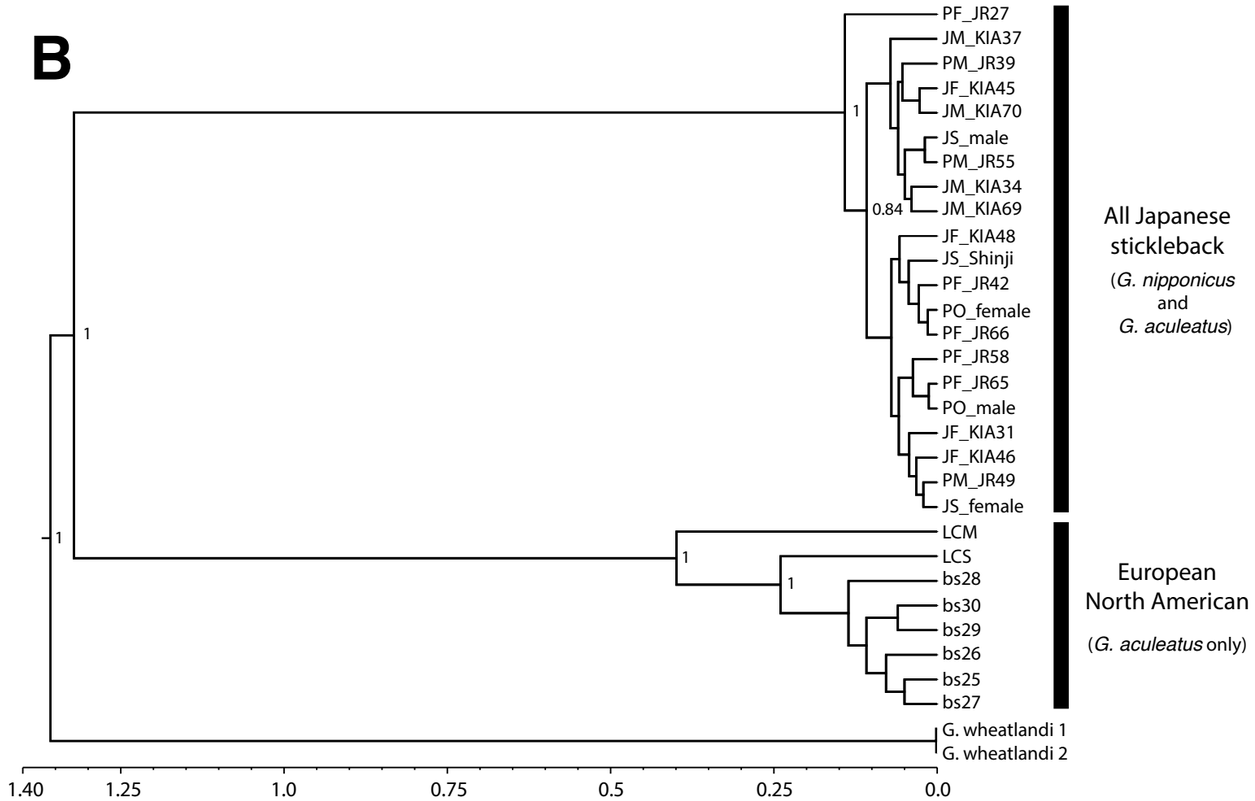
1454 **S3 Fig. Bootstrapped PSMC curves for 26 resequenced individuals.**

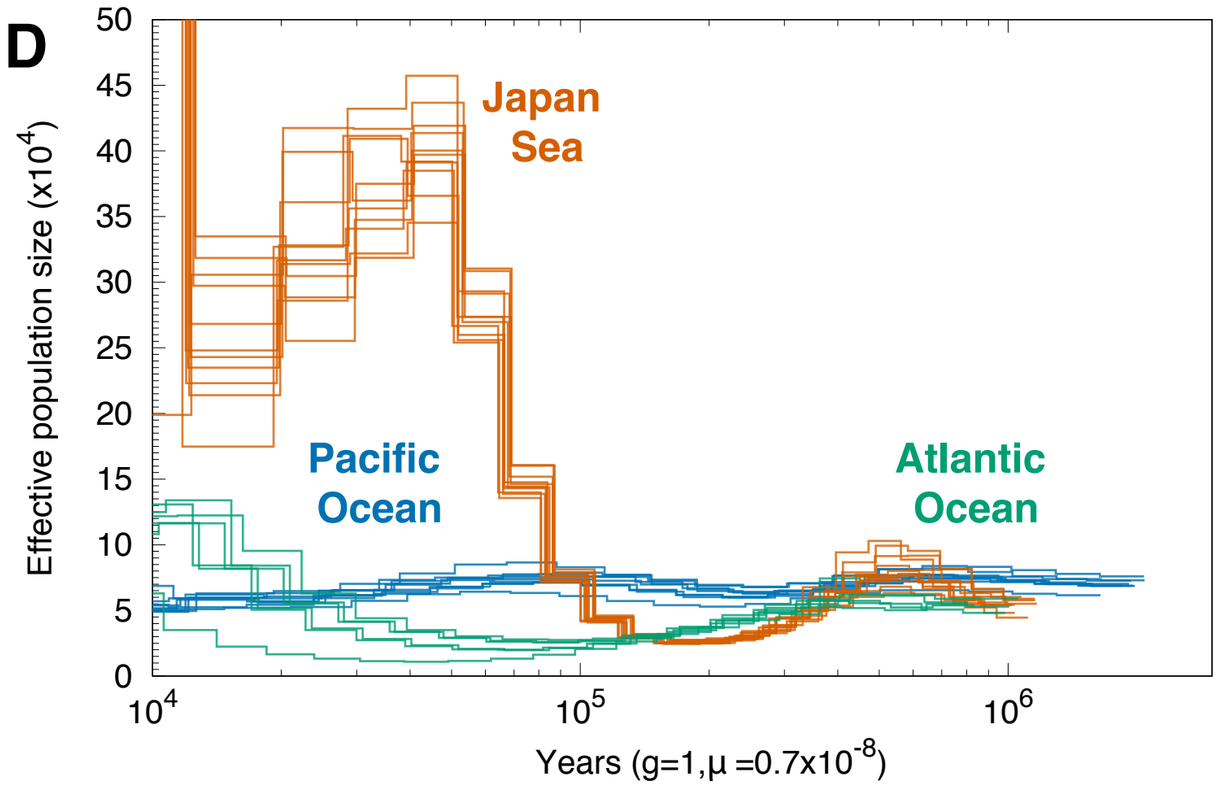
1455 **S4 Fig. PSMC profile for all 26 individuals and an additional Eastern Pacific**
1456 **individual from Little Campbell River, Canada.**
1457 **S5. Fig. Principal component analysis on RAD-seq data from 295**
1458 **individuals from across the distribution of all three lineages.** The arrow
1459 indicates the presence of an admixed individual occurring in the Akkeshi system.
1460 **S6 Fig. STRUCTURE analysis on RAD-seq data from Japanese populations**.
1461 Analysis with K = 2 clusters (A), which is supported by likelihood analysis (B),
1462 showed the presence of admixed individuals in the Akkeshi system.
1463 **S7 Fig. Genome-wide F_{ST} measured in non-overlapping 50 kb windows for**
1464 **allopatric and sympatric between and within species comparisons.**
1465 **S8 Fig. Genome-wide d_{XY} measured in non-overlapping 50 kb windows for**
1466 **allopatric and sympatric between and within species comparisons.**
1467 **S9 Fig. Negative association between f_d and G_{MIN} in sympatric (JS vs PO) and**
1468 **allopatric (JS v AT) between species comparisons.**
1469 **S10 Fig. Genome-wide G_{MIN} for sympatric between species comparisons.**
1470 Black line represents 50 kb non-overlapping window G_{MIN} signature. Points
1471 represent 10 kb windows; grey points are non-valley windows, blue points are
1472 valley windows identified by Hidden Markov Model algorithm.
1473 **S11 Fig. Absolute divergence (d_{XY}) is lower and f_d is higher in G_{MIN} valleys**
1474 **compared to the genome-wide background.**
1475 **S12 Fig. Genome-wide f_d for sympatric between species comparisons.** Black
1476 line represents 50 kb non-overlapping window f_d signature. Points represent 10
1477 kb windows; grey points are non-valley windows, blue points are peak windows
1478 identified by Hidden Markov Model algorithm.
1479 **S13 Fig. Comparison of d_{XY} between f_d peaks and non-peaks.**

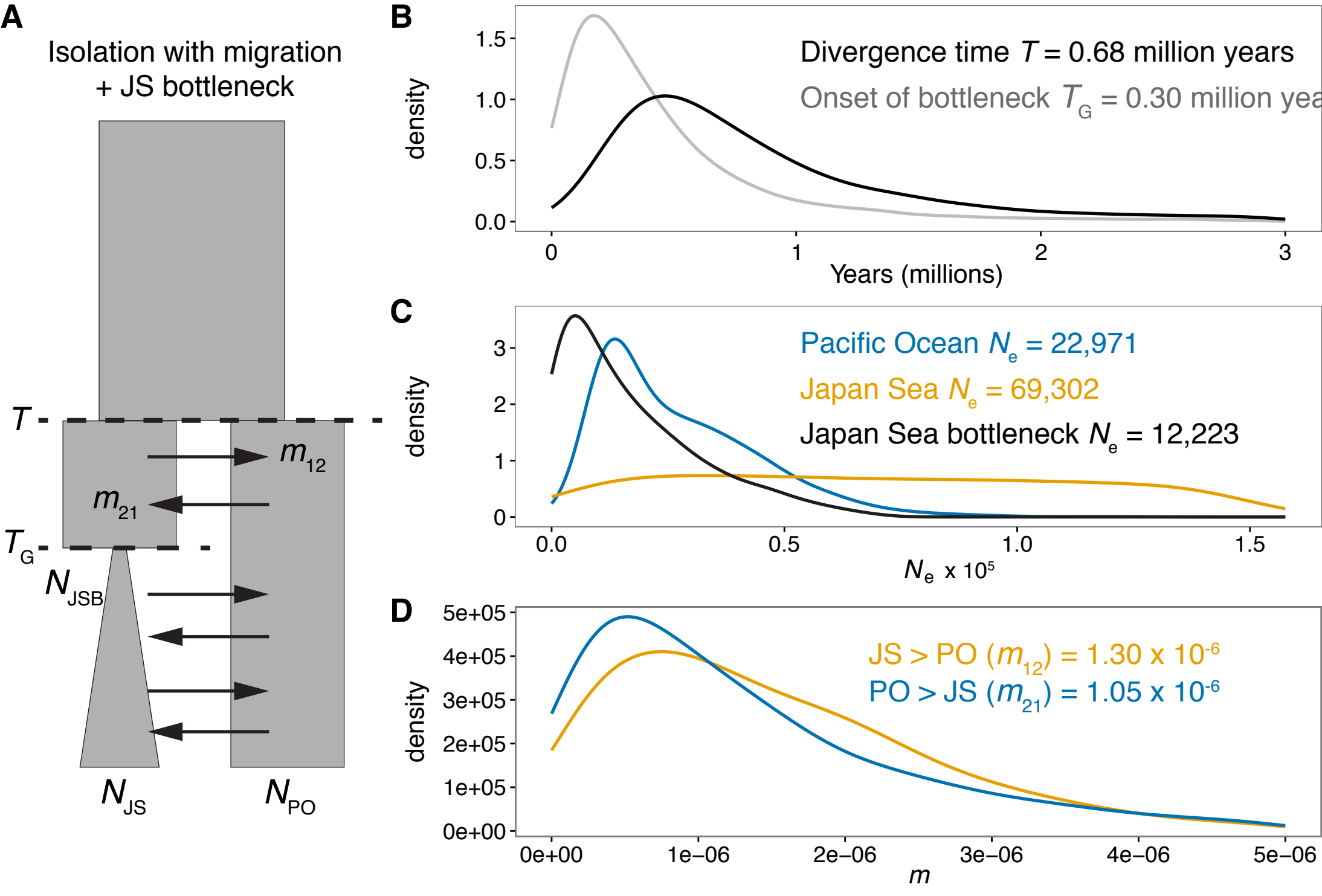
1480 **S14 Fig. Analysis of partitioned *D* statistics.** (A) Boxplots comparing
1481 partitioned *D* statistics between f_d peaks and the autosomal background. Dashed
1482 line at zero indicates a balance between allele patterns indicative of incomplete
1483 lineage sorting. In (B), P1 = Atlantic Ocean (AT), P2 = Pacific Ocean (PO), $P3_1$ =
1484 sympatric Japan Sea (JS_S), $P3_2$ = allopatric Japan Sea (JS_A), and O = *G. wheatlandi*
1485 (WT). In (C), P12 and P3 were swapped. D_1 measures asymmetry between P1
1486 and P2 where the derived allele B is present in $P3_1$ but not $P3_2$, D_2 measures
1487 where allele B is present in $P3_2$ but not $P3_1$, and D_{12} measures where the derived
1488 allele is shared by both $P3_1$ and $P3_2$. If we assume that the derived allele B
1489 occurred at the ancestor of P3, D12 indicates introgression from P3 to P2. See
1490 [56] for a more detailed explanation of these statistics.
1491 **S15 Fig. D_{FOIL} statistics.** In this statistics, we assume that the divergence time
1492 between P_1 and P_2 is younger than that between P_3 and P_4 and infer all possible
1493 introgressions including ancestral introgression involving P_{12}. The number of
1494 loci (100 kb-window) that show statistically significant introgression are shown.
1495 See [57] for a more detailed explanation of these statistics.
1496 **S16 Fig. The relationship between introgression measured as G_{MIN} and f_d**
1497 **and $\log_{10}$ recombination rate.**
1498 **S17 Fig. Venn diagram showing overlap between genes occurring in**
1499 **introgressed regions identified using different measures.**
1500 **Supplementary methods - ABC analysis.**

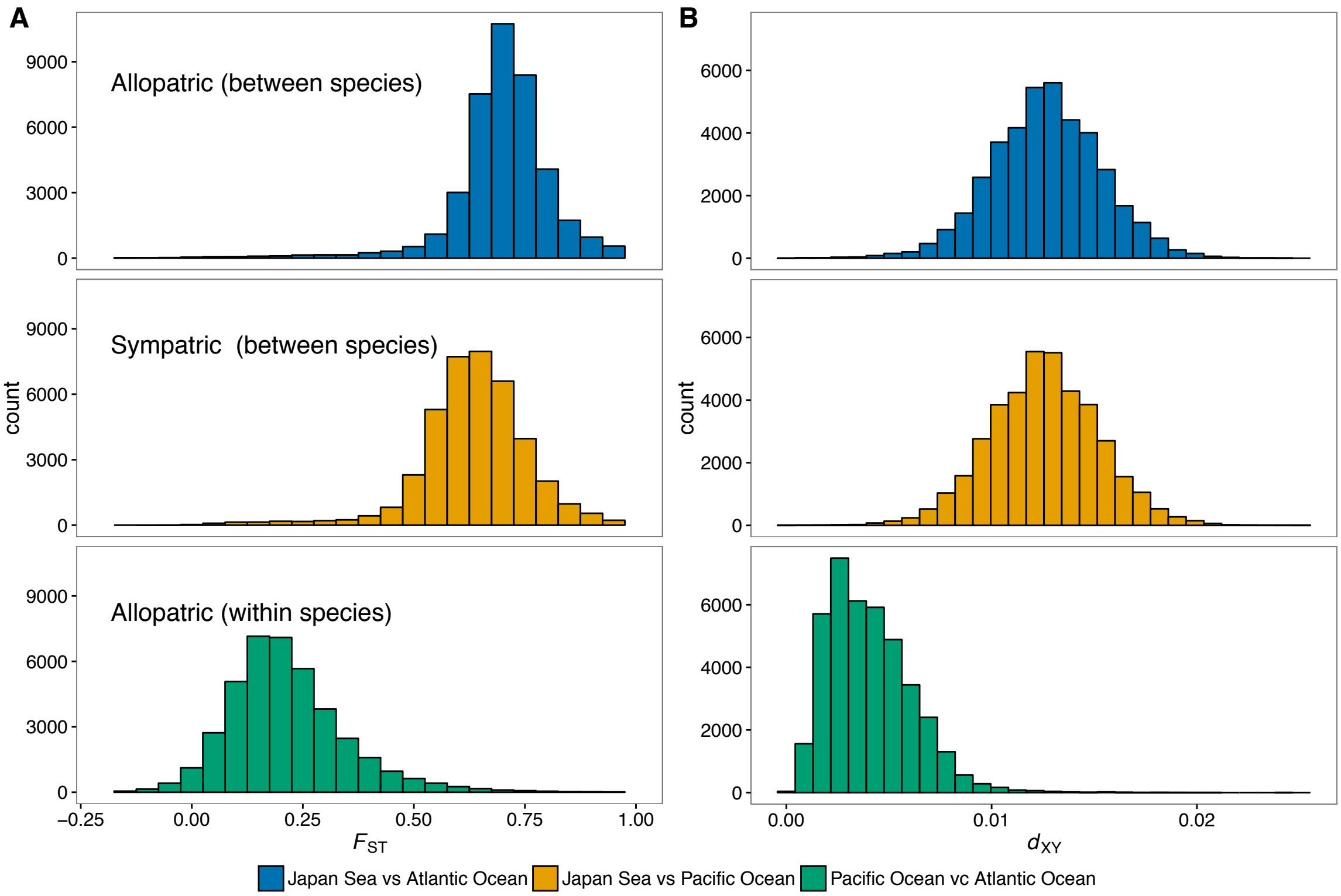

A
bs28
bs25
bs27
bs26
bs30
bs29
PF_JR65
PF_JR27
PM_JR49
PF_JR42
PO_male
PO_female
PF_JR58
PM_JR39
PM_JR55
PF_JR66
JF_KIA48
JS_male
JM_KIA69
JM_KIA37
JS_female
JF_KIA45
JF_KIA46
JF_KIA31
JM_KIA70
JM_KIA34
G. wheatlandi
Atlantic Ocean
(G. aculeatus)
Pacific Ocean
(G. aculeatus)
Japan Sea
(G. nipponicus)

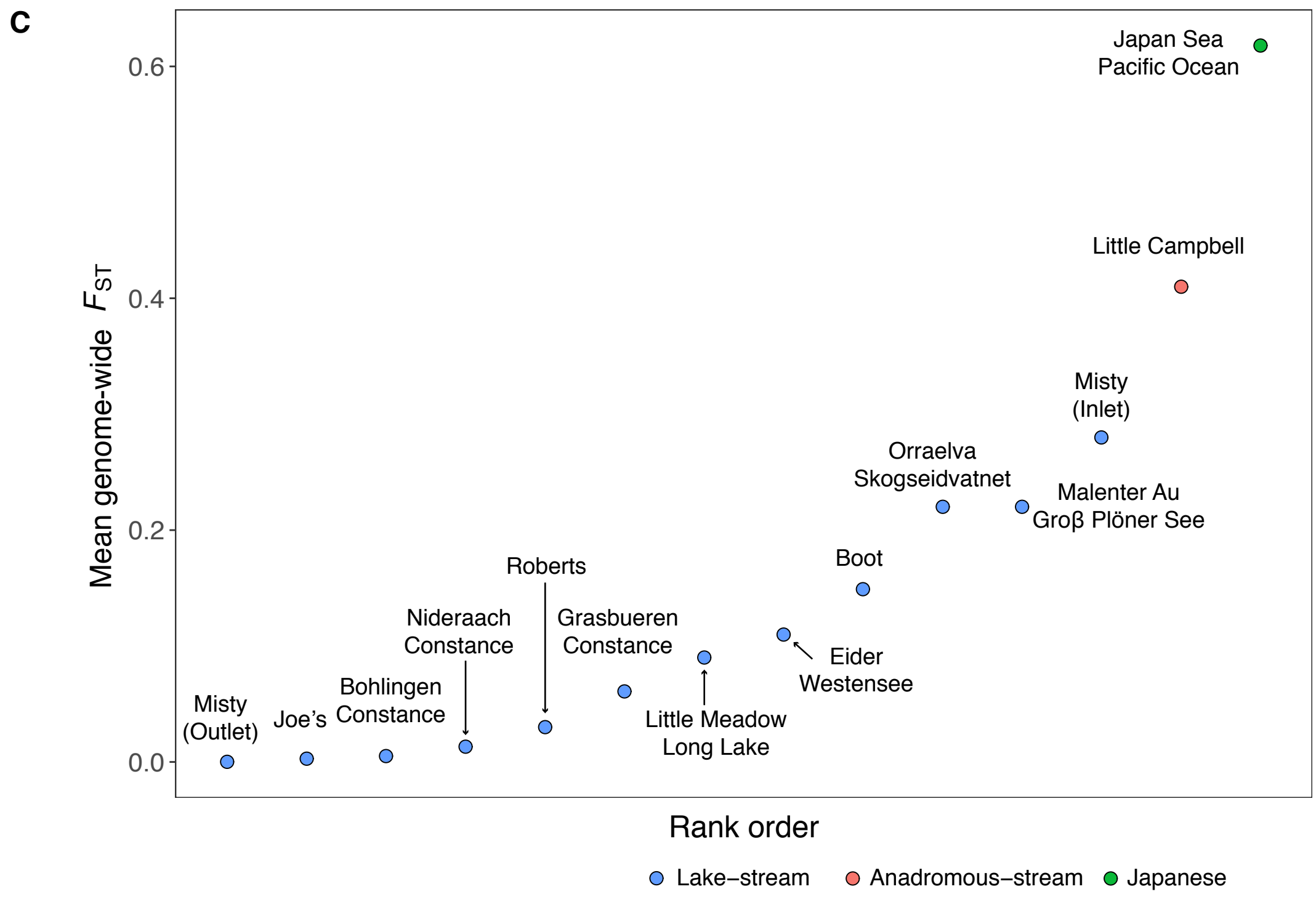

C
Akkeshi system
G. nipponicus
G. aculeatus
Sea of Japan
Pacific Ocean
N

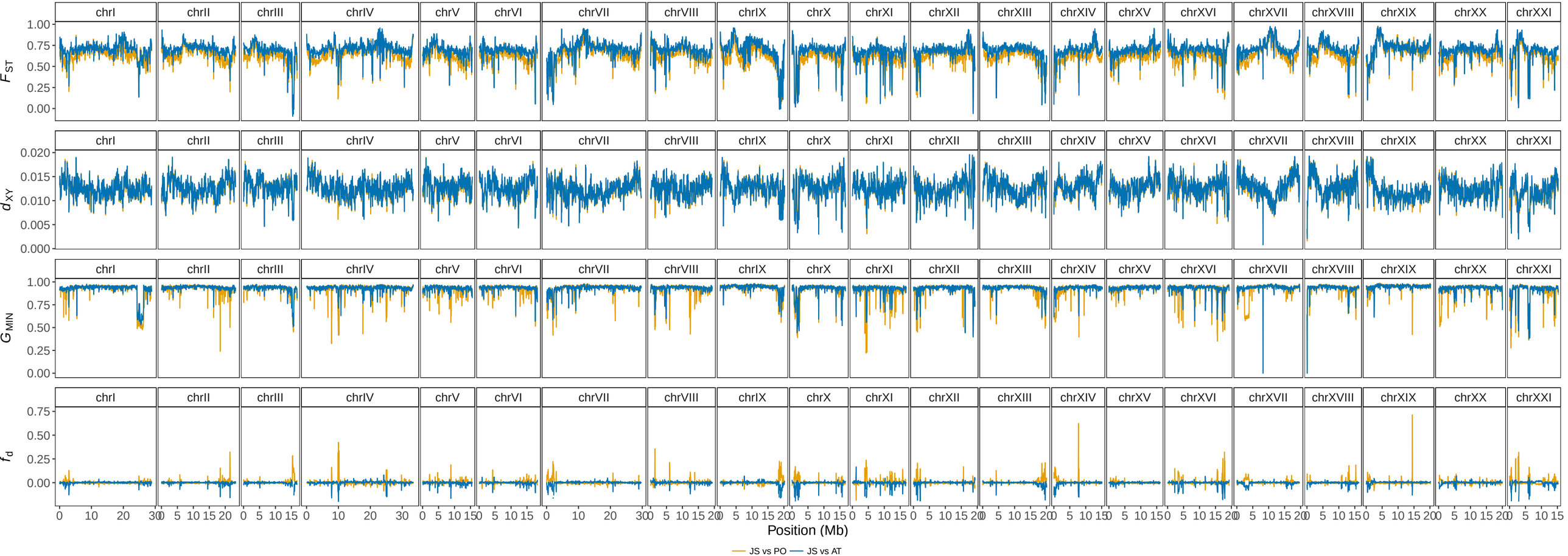

B
PF_JR27
JM_KIA37
PM_JR39
JF_KIA45
JM_KIA70
JS_male
PM_JR55
JM_KIA34
JM_KIA69
JF_KIA48
JS_Shinji
PF_JR42
PO_female
PF_JR66
PF_JR58
PF_JR65
PO_male
JF_KIA31
JF_KIA46
PM_JR49
JS_female
LCM
LCS
bs28
bs30
bs29
bs26
bs25
bs27
G. wheatlandi 1
G. wheatlandi 2
1
0.84
1
1
1
1
All Japanese stickleback
(G. nipponicus and G. aculeatus)
European North American
(G. aculeatus only)
1.40
1.25
1.0
0.75
0.5
0.25
0.0

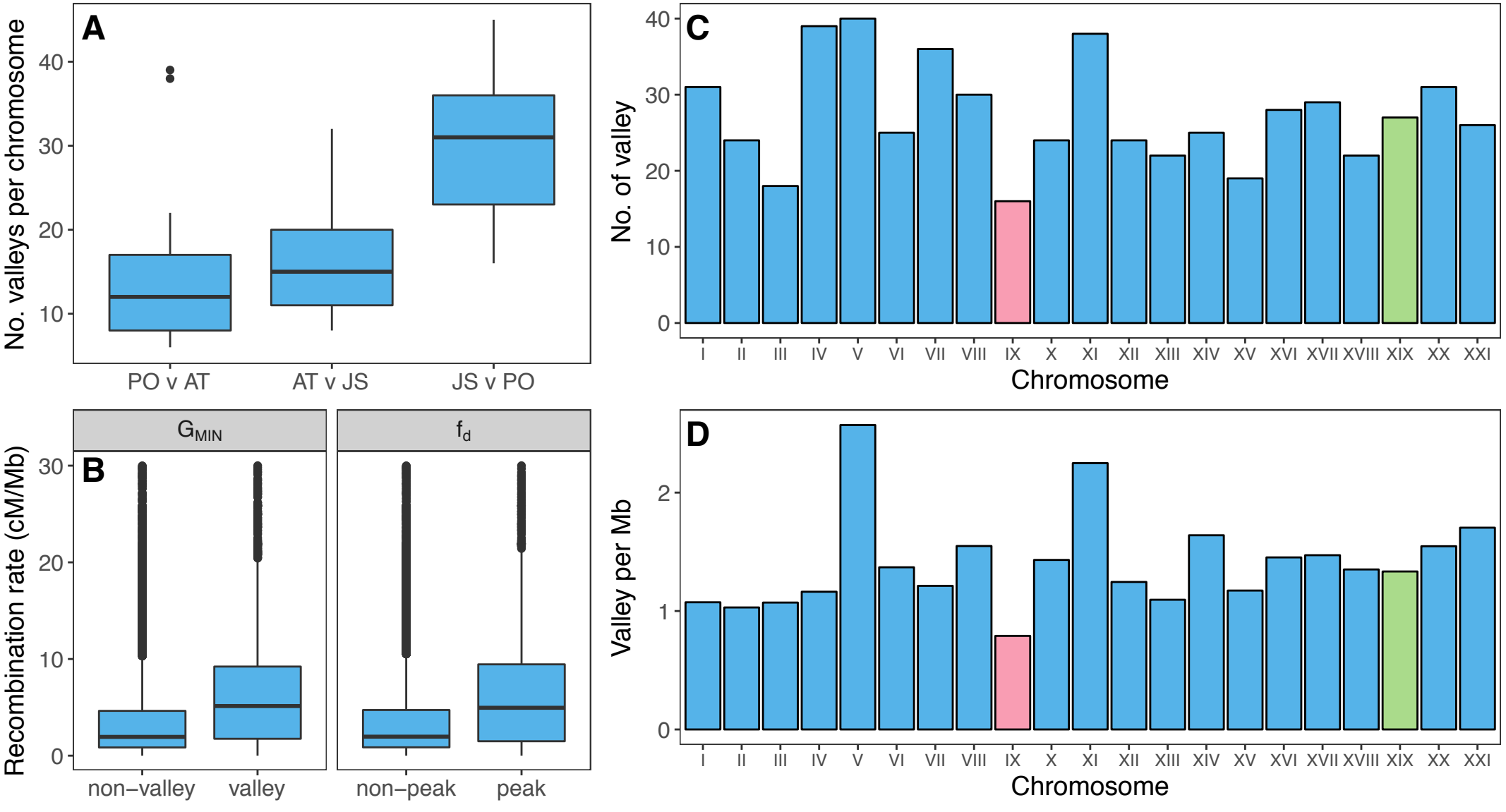

D
Effective population size (x10^4)
50
45
40
35
30
25
20
15
10
5
0
Japan Sea
Pacific Ocean
Atlantic Ocean
10^4
10^5
10^6
Years (g=1,μ =0.7x10^-8)


A
Isolation with migration
+ JS bottleneck
T
m12
m21
TG
NJSB
NJS
NPO
B
Divergence time T = 0.68 million years
Onset of bottleneck TG = 0.30 million yea
density
Years (millions)
C
Pacific Ocean Ne = 22,971
Japan Sea Ne = 69,302
Japan Sea bottleneck Ne = 12,223
density
Ne x 10^5
D
JS > PO (m12) = 1.30 x 10^-6
PO > JS (m21) = 1.05 x 10^-6
density
m


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
Groß Plöner See
Boot
Roberts
Nideraach
Constance
Grasbueren
Constance
Eider
Westensee
Bohlingen
Constance
Misty
(Outlet)
Joe's
Little Meadow
Long Lake
Lake–stream
Anadromous–stream
Japanese


chrI
chrII
chrIII
chrIV
chrV
chrVI
chrVII
chrVIII
chrIX
chrX
chrXI
chrXII
chrXIII
chrXIV
chrXV
chrXVI
chrXVII
chrXVIII
chrXIX
chrXX
chrXXI
F_{ST}
d_{XY}
G_{MIN}
f_d
Position (Mb)
JS vs PO
JS vs AT


A
No. valleys per chromosome
40
30
20
10
PO v AT
AT v JS
JS v PO
B
G_{MIN}
f_d
Recombination rate (cM/Mb)
30
20
10
0
non−valley
valley
non−peak
peak
C
No. of valley
40
30
20
10
0
I
II
III
IV
V
VI
VII
VIII
IX
X
XI
XII
XIII
XIV
XV
XVI
XVII
XVIII
XIX
XX
XXI
Chromosome
D
Valley per Mb
2
1
0
I
II
III
IV
V
VI
VII
VIII
IX
X
XI
XII
XIII
XIV
XV
XVI
XVII
XVIII
XIX
XX
XXI
Chromosome