Balancing Selection in Structured Populations of Three-Spined Sticklebacks

Doko-Miles Thorburn (✉ d.m.j.thorburn@qmul.ac.uk)

Queen Mary University of London <https://orcid.org/0000-0002-0120-8829>

Kostas Sagonas

Queen Mary University of London

Tobias Lenz

Max Planck Institute for Evolutionary Biology

Frederic Chain

University of Hamburg

Philine Feulner

EAWAG Swiss Federal Institute of Aquatic Science and Technology

Erich Bornberg-Bauer

Institute for Evolution and Biodiversity

Thorsten Reusch

GEOMAR Helmholtz Centre for Ocean Research Kiel

Manfred Milinski

Max-Planck-Institute for Evolutionary Biology

Christophe Eizaguirre

Queen Mary University of London

Article

Keywords: balancing selection, *Gasterosteus aculeatus*, demography, genome scan

Posted Date: November 30th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-1092984/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Balancing Selection in Structured Populations of Three-Spined Sticklebacks

Doko-Miles J. Thorburn^{1,2+}, Kostas Sagonas^{1,3}, Tobias L. Lenz^{4,5}, Frederic J.J. Chain⁶, Philine G.D. Feulner^{7,8}, Erich Bornberg-Bauer⁹, Thorsten BH Reusch¹⁰, Manfred Milinski¹¹, Christophe Eizaguirre¹

¹School of Biological and Chemical Sciences, Queen Mary University of London, London, UK.

²Department of Life Sciences, Imperial College London, London, UK.

³Faculty of Biosciences and Aquaculture, Nord University, Bodø, Norway.

⁴Research Group for Evolutionary Immunogenomics, Max Planck Institute for Evolutionary Biology, Plön, Germany.

⁵Research Unit for Evolutionary Immunogenomics, Department of Biology, University of Hamburg, Hamburg, Germany.

⁶Department of Biological Sciences, University of Massachusetts Lowell, USA.

⁷Department of Fish Ecology and Evolution, Centre of Ecology, Evolution and Biogeochemistry, EAWAG Swiss Federal Institute of Aquatic Science and Technology, Kastanienbaum, Switzerland.

⁸Division of Aquatic Ecology and Evolution, Institute of Ecology and Evolution, University of Bern, Baltzerstrasse 6, 3012 Bern, Switzerland

⁹Evolutionary Bioinformatics, Institute for Evolution and Biodiversity, Westfälische Wilhelms University, Münster, Germany.

¹⁰Marine Evolutionary Ecology, GEOMAR Helmholtz Centre for Ocean Research Kiel, Kiel, Germany.

¹¹Department of Evolutionary Ecology, Max Planck Institute for Evolutionary Biology, Plön, Germany.

⁺Corresponding author: Thorburn D. M. J. (d.m.j.thorburn@qmul.ac.uk)

Abstract

Balancing selection describes evolutionary processes that maintain genetic diversity. To date, the number of impacted genes and underlying biological functions remain elusive. Using 60 three-spined stickleback genomes (*Gasterosteus aculeatus*) from five recently diverged lake-river population-pairs, we performed genome-wide scans across two levels of organization: population-pairs and populations. We overlapped Tajima's D and Watterson's estimator metrics and verified signals with additional summary statistics, and evaluated alternative explanations: neutral evolution, population structure, associated overdominance, or demographic change. Candidate windows exhibiting signals of balancing selection spanned 2.31% (population-pair) and 3.10% (population) of the autosomes. These candidate windows had extended linkage disequilibrium and were enriched in intergenic and non-synonymous SNPs. We identified 715 (population-pair) and 1,010 (population) candidate genes under balancing selection. Importantly, using conservative thresholds, we found a small proportion of candidate genes overlapped with highly differentiated loci or regions of potential associated overdominance. There was little evidence of confounding effects originating from demographic change. Overall, candidate genes under balancing selection were associated with functions related to interactions with the environment (olfaction and receptor signalling pathways). Our results demonstrate selection that maintains standing genetic variation is common and evolves in response to local environmental pressures, playing an important role in adaptation.

Keywords: balancing selection, *Gasterosteus aculeatus*, demography, genome scan

Background

Understanding the evolutionary processes that maintain genetic diversity is a central goal in evolutionary biology^{1–3}. One such process is balancing selection⁴, characterised by the maintenance of alleles at low to intermediate frequencies that may later become positively selected, facilitating evolutionary responses from standing genetic variation^{5–7}. Balancing selection results from several non-mutually exclusive mechanisms, such as heterozygote advantage (and overdominance), negative frequency-dependent selection, sexually antagonistic selection, antagonistic pleiotropy, or temporally and spatially heterogeneous selection^{4,8–10}. Notably, all forms of balancing selection can act across time and space in response to habitat heterogeneity, affecting allele frequencies within and among populations^{11–16}.

The majority of characterized genes under balancing selection stem from *a priori* candidate gene approaches in host-parasite coevolution studies, focusing on immunity genes^{5,17,18}. Classical examples include the resistance *R*-genes in plants¹⁹, the major histocompatibility complex (MHC) gene region in jawed vertebrates (e.g. Takahata and Nei 1990; Penn et al. 2002; Arora et al. 2020), and invertebrate antimicrobial genes^{23,24}. There are also, however limited, other genes known to evolve under balancing selection such as the self-incompatibility locus in plants^{25,26}, the complementary sex determination in honey bees²⁷, the migration-related *PGI* gene in Glanville fritillary butterfly¹⁴, the ABO blood group system in primates²⁸, and the sexually-antagonistic *VGLL3* locus in salmon²⁹. Together those examples suggest that balancing selection may affect diverse functions.

Whilst genome-wide scans have expanded our understanding of how positive and purifying selection affect genes' evolution^{30–32}, there are only a few genome scans of balancing selection^{24,33–36}. Consequently, it is unclear how many genes are affected, and which functions they are associated with. Studies report a wide range of genes under balancing selection: 0, 60, and 1,404–1,616 candidate genes in humans^{33,37,38}, 68 genes in *Ralstonia solanacearum*³⁴, 99 genes in *Staphylococcus aureus*³⁵, and 183 genes in *Drosophila melanogaster*²⁴. The disparity in the number of genes evolving under balancing selection in humans for instance is likely due to the challenges associated with its detection, such as verifying signals with multiple lines of evidence and identifying the best organization level to be investigated, e.g. population, region, or species^{9,39}.

Theoretical signatures of balancing selection include increased nucleotide diversity, excess of polymorphisms at intermediate frequencies, excesses of nonsynonymous polymorphisms, and extended linkage disequilibrium (LD,^{4,9}). However, tests for the detection of these signatures can be confounded by other types of natural selection, associated overdominance

(AOD), gene flow, and rapid demographic change, affecting their application and interpretation^{9,40,41}. For example, Non-Central Deviation (NCD), a novel statistic that captures departures from allele frequencies expected under balancing selection, is confounded by incomplete sweeps of positive selection, demography, and AOD³³. AOD can generate genetic diversity in linked neutral or deleterious loci when recombination rates are low, and is confounded by balancing selection^{40,41}. Identifying regions putatively evolving under AOD can be achieved by overlapping regions of low recombination with outlier loci to determine the likelihood the observed signals are either balancing selection or AOD. Tajima's D , can detect excesses of polymorphism at intermediate frequencies, and is confounded by demographic histories and gene flow^{9,42}. The McDonald-Kreitman test, which detects an excess of nonsynonymous polymorphism amongst species, is confounded by background selection and requires coding sequences rather than regulatory elements^{9,43}. Although, the McDonald-Kreitman test has also been applied to non-coding DNA, these applications are constrained by availability and quality of annotations⁴⁴. Altogether, it has become essential that scans for balancing selection use multiple lines of evidence and are conducted in a clear demographic and population structure framework.

Genomic signature of balancing selection that contributes to local adaptation and fuels population-specific evolution should be characterised by a high proportion of population-specific candidate genes^{14,45,46}. Genetically and functionally resolved examples of balancing selection contributing to local adaptation are rare, but are known in the Low germination stimulant 1 (*LGS1*) gene in *Sorghum bicolor*⁴⁷ and the MHC gene complex in *Gasterosteus aculeatus*⁴⁸. In both cases, the mechanisms of balancing selection work over short ecological timescales, suggesting spatiotemporal heterogeneous selection or negative-frequency dependent selection are likely the dominant mechanisms. Noteworthy, using replicated populations further enables investigating whether parallel genetic patterns of balancing selection exist in addition to adaptation to environmental heterogeneity within a location-specific context.

Since the last glaciation (ca. 11 kya), the three-spined stickleback (*Gasterosteus aculeatus*) has repeatedly colonised and adapted to freshwater habitats across the northern hemisphere, forming ecotypes that differ in morphology, physiology, and behaviours^{49–55}. Parapatric lake-river population-pairs show evidence of parallel evolution^{53,55–57}, but mostly demonstrate that population differentiation stems from population-specific selection and consequently local adaptation^{54,58,59}. Here, we performed genome scans across a structured sampling design with ten populations from five lake-river population-pairs using classical neutrality tests (Tajima's D and Watterson θ ;^{42,60}). We verified the call for balancing selection by determining empirical distributions of nucleotide diversity and the Non-Central Deviation statistic (NCD;³³), before

simulating neutral expectations emerging from population structure. Additionally, we tested for functional enrichment among genes in candidate regions, determined their linkage disequilibrium, and population differentiation. To evaluate the roles of alternative sources of genetic variation, we also scanned for AOD stemming from strong linkage disequilibrium and low recombination rates. All tests were established within an appropriate demographic framework by reconstructing past effective population sizes using MSMC⁶¹.

Results

Primary Genome Scans

We used whole-genome resequencing data of 60 *G. aculeatus* individuals, representing 10 populations from 5 recently diverged freshwater parapatric lake-river population pairs (details about sampling, library preparation, and sequencing are in ⁵⁹). The five populations-pairs were sampled from connected lakes (_L) and rivers (_R) in Canada (CA), the United States (US), Norway (NO), and two German sites (G1 and G2). Whereas previous analyses of these populations have included genome scans for increased population differentiation ⁵⁹, here we focus on scans of increased polymorphism. Sequences were mapped to the *G. aculeatus* reference genome (Broad/gasAcu1;⁶²). A total of 9,916,056 SNPs reaching an average genome wide depth of coverage of 22 (range 11-44; supplementary table S1) were identified. The proportion of missing data in sites genotyped across the dataset ranged from $5.03 \times 10^{-3} \pm 2.33 \times 10^{-4}$ (US_R; genome-wide mean $\pm$ standard error; SE) to $0.010 \pm 4.07 \times 10^{-4}$ (G2_R). Genomic windows with more than 1% of missing data were masked for downstream analyses.

To identify signatures of balancing selection, we generated an empirical distribution of Tajima's D (T_D ; Tajima 1989) and Watterson's estimator (θ_w ; Watterson 1975). We used overlapping scanning windows of 7.5 kb and a step size of 1.5 kb, ensuring considerable overlap between neighbouring windows to detect narrow signatures consistent with balancing selection. We scanned across two organization levels: regional population-pair (pooling lake and river fish from the same watershed, $N=12$ per pair), and population ($N=6$ per population). Autosomal averages including the unmapped chromosome for T_D ranged from -0.112 ± 0.001 (US) to 0.567 ± 0.002 (CA) at the population-pair level, and from -0.143 ± 0.001 (NO_L) to 0.506 ± 0.002 (G1_R) at the population level. Similarly, autosomal averages for θ_w ranged from $1.81 \times 10^{-3} \pm 1.75 \times 10^{-6}$ (NO) to $3.21 \times 10^{-3} \pm 2.99 \times 10^{-6}$ (US) at the population-pair level, and $1.24 \times 10^{-3} \pm 1.59 \times 10^{-6}$ (NO_R) to $3.23 \times 10^{-3} \pm 3.24 \times 10^{-6}$ (US_L) at the population level (supplementary table S2).

Distribution of Candidate Windows and Candidate SNPs

We defined genomic windows with signatures of balancing selection (hereafter candidate windows) as windows with both T_D above 2.0, and θ_w two standard deviations above the

population-pair (PP) mean, or population (Pop) mean (Croze et al. 2017; supplementary table S2). Combining both indices, we detected 5,853 (PP) and 6,412 (Pop) overlapping candidate windows (i.e., consecutive candidate windows or those identified in more than 1 population-pair or population), spanning a total of 10.3Mb (PP; 2.31%), and 12.7Mb (Pop; 3.10%) of the genome (fig. 1, supplementary fig. S1, table 1).

To test whether balancing selection disproportionately affects particular types of functional sites, we annotated SNP locations in candidate windows. Empirical P values were obtained by comparing the proportion of candidate SNPs in each category (exon, intron, intergenic, splice site, or regulatory) to the genomic baseline using a circular permutation approach (supplementary table S3). Across both levels of analysis, when comparing to the genomic baseline we found enrichment of each SNP category in candidate windows, with the exception of splice sites ($P \leq 0.001$; fig. 2 and supplementary fig. S2). Exonic SNPs were enriched in candidate windows in 1 population-pair (NO) and 3 populations (G1_L, G2_L, NO_R; $P \leq 0.014$). Intronic SNP enrichment was observed 1 population (CA_R; $P = 0.004$). Regulatory SNPs were enriched in 1 population (NO_L; $P = 0.001$). Strikingly, we found enrichment of nonsynonymous SNPs, which directly affect the protein sequence, in 3 population-pairs (G1, NO, US) and 4 out of 10 populations (G1_L, G1_R, G2_L, NO_R; $P \leq 0.013$). Whilst these analyses are limited by the quality of annotations, especially for regulatory sites, targets of balancing selection are different in different populations and disproportionately affect protein structure across both levels of analysis.

Distribution of Genes and GO Enrichment

Within the candidate genomic regions, we identified a total of 971 (PP) and 1,405 (Pop) candidate genes, of which 715 (PP; 3.44% of *G. aculeatus* protein coding genes) and 1,010 (Pop; 8.46%) had unique Ensembl gene identifications (supplementary table S4). Candidate windows and genes were not randomly distributed across the genome, with 66.43% (PP) and 60.04% (Pop) candidate windows, and 42.12% (PP), and 41.85% (Pop) candidate genes, located within three chromosomes: IV, VII, and XXI (X^2 tests are reported in table 2). Overall, these three chromosomes contain 13.86% of *G. aculeatus* genes.

The majority of candidate genes were private: 504 out of 715 (PP; 70.5%) genes were found only in a single population-pair (fig. 3A), while 738 out of 1,010 (Pop; 73.1%) were found in a single population (fig. 3B). 27.38% (PP) and 27.37% (Pop) of these private candidate genes were located on chromosomes IV and VII, respectively, and overall were not randomly distributed across the genome (PP, $X^2 = 374.1$, $df = 20$, $P < 0.001$; Pop, $X^2 = 522.1$, $df = 20$, $P < 0.001$; fig. 3C). Balancing selection also affected the same genes across multiple populations, hereafter called *parallel candidate genes*, with 211 candidates shared among two or more population-pairs (fig. 3A) and 272 among populations (fig. 3B). Additionally, 450

candidate genes were identified in both levels of analysis. There were several candidates identified in 3 or more population-pairs and in at least 6 populations: *Plxnc1*, *VMA21*, *FYCO1A*, *XCR1A*, *NXPE3*, and an unnamed gene (*ENSGACG00000011738*; fig 4A-C, supplementary fig. S3). Interesting, *FYCO1A* and *XCR1A* may have hitchhiked on the exceptionally strong signal of selection in *ENSGACG00000011738* based on its close proximity (fig. 4A).

Subsequently, we performed a gene ontology (GO) enrichment analysis on sets of candidate genes. 582 of 715 (PP; 81.40%), and 808 of 1,010 (Pop; 80%) *G. aculeatus* genes were assigned *Danio rerio* orthologs. There were no significantly enriched GO terms in the population-pair analysis (supplementary table S5). At the population level, we observed significant enrichment in “Sensory perception of smell” (FDR < 0.001), and “Glycosaminoglycan (GAG) catabolic process” (FDR = 0.031; supplementary table S6). To better understand the distribution of those enrichment terms, we split enrichment of candidate genes at the population level for those displaying i) signals of parallel balancing selection, ii) private balancing selection (found only in one population), and iii) candidate genes showing copy number variation (CNV, Chain et al. 2014). Among 272 candidate genes with signals of parallel balancing selection, we assigned 218 orthologs. We found one enriched function, “Sensory perception of smell” (FDR = 0.014; supplementary table S7), which is exclusively composed of genes from the Olfactory Receptor superfamily. 590 of 738 private candidate genes were assigned *D. rerio* orthologs, and we observed enrichment in “GAG catabolic process” (FDR = 0.013; supplementary table S8), including 4 of 6 annotated GAG catabolism genes in the genome (*SGSH*, *LYGL1*, *CEMIP2*, and *PGLYRP2*). Finally, 51 of 84 copy number variant (CNV) candidate genes were assigned orthologs, and enrichment was observed in “G protein-coupled receptor signalling pathway” (FDR = 0.002) and “Sensory perception of smell” (FDR < 0.001; supplementary table S9). Overall, our GO enrichment analyses reveal that balancing selection is associated with several gene functions, which are largely associated with the capacity of individuals to interact with their environment, likely as a result of environmental heterogeneity.

Performance of Genome Scans

In order to confirm the signatures of balancing selection, we compared nucleotide diversity (π) and the non-central deviation statistic (NCD1;³³) in candidate windows to the genomic baseline. NCD1 detects shifts in the site frequency spectrum from equilibrium frequencies to those expected under balancing selection, and has previously been used to independently validate signals of balancing selection⁶³. We compared π and NCD1 distributions in candidate windows with a null distribution estimated from randomly sampling non-candidate windows. As expected, autosomal estimates of π were higher in candidate windows than for the genomic baseline (LMM; PP, $5.92 \times 10^{-3} \pm 5.08 \times 10^{-5}$ (estimate $\pm$ SE), $F_{2031023} = 13606$, $P < 0.001$; Pop,

5.70×10⁻³ ± 4.49×10⁻⁵, $F_{1431419} = 16153$, $P < 0.001$). Similarly, autosomal deviations from the NCD1 target frequency were smaller than the genomic baseline (LMM; PP, 0.119 ± 1.29×10⁻³ (estimate ± SE), $F_{1903840} = 7505.2$, $P < 0.001$; Pop, 0.133 ± 1.83×10⁻³, $F_{1297954} = 5261.8$, $P < 0.001$). High π and low NCD1 are expected characteristics of balancing selection, hence confirming our genome scan correctly identified genomic candidate regions and their associated biological functions.

Linkage Disequilibrium, Associated Overdominance, and Differentiation

We tested for the prediction that linkage disequilibrium (LD) should be extended around sites under balancing selection. The unmapped chromosome was excluded from this analysis. Candidate windows exhibited significantly extended LD when compared to the baseline, and LD was detectable in all 10kb windows up to 150kb away (LMM, 0kb-10kb, $F_{5949} = 2414.5$, $FDR < 0.001$; 140kb-150kb, $F_{4999} = 24.7$, $FDR < 0.001$; fig. 5A, supplementary table S10). We also detected a positive correlation between the size of the genomic regions under balancing selection and the strength of LD in all 10kb windows up to 90kb (LMM, 0kb-10kb, $F_{25382} = 88.2$, $FDR < 0.001$; 80kb-90kb, $F_{23401} = 8.43$, $FDR = 0.011$; fig. 5B, supplementary table S10).

We also investigated whether private candidate genes had more extended LD than i) parallel candidate genes, ii) non-candidate genes that were candidates in other populations (hereafter quasi-candidates), and iii) the genomic baseline. Here, we hypothesized that private candidate genes should be associated with extended LD only in those populations, and LD should be comparable in size to those of parallel genes under balancing selection. The split of private and parallel candidate genes further enables addressing evolutionary timing of selection, with private candidate genes hypothesized to be under more recent balancing selection and therefore having the highest mean R^2 . In general, mean R^2 values were highest in private candidate genes in 9 of 20 10kb bins when compared to parallel candidates and quasi-candidates, all of which had significantly higher R^2 compared to the genomic baseline (LMM; $F_{177274} = 1930.4$, $P < 0.001$; supplementary table S11 and fig. S4). As hypothesized, this result suggests private candidate genes are likely evolving in response to population-specific selection and evolving under more recent balancing selection.

Because extended LD might come with confounding factors, we investigated the association between balancing selection, associated overdominance (AOD), and population differentiation. Genetic diversity in regions of low recombination can be generated by AOD, mimicking signals of balancing selection⁴⁰. Additionally, theory predicts that gene flow will obscure signals of natural selection unless islands of differentiation are formed around regions under balancing selection^{9,64}. Hence, we investigated whether candidate windows were associated with either regions of low recombination or islands of differentiation. Regions of low recombination were defined as the bottom 10% of the recombination rate distribution, and

F_{ST} outliers were defined as the top 10% of the empirical F_{ST} distributions. By determining the proportion of candidate windows that overlap with the conservatively estimated recombination rate or F_{ST} outliers, we can identify the proportion of candidate windows that can only realistically be explained by natural selection, and not AOD or population structure. Specifically, a total of 47.96% (2,807) and 35.87% (2,300) of candidate windows and 19.05% (185 total; 113 unique) and 18.65% (262 total; 172 unique) of candidate genes overlapped with low recombination regions at the population-pair and population levels, respectively. However, only 2 population-pairs (G1, G2) and 3 populations, all being from river origin (G1_R, G2_R, NO_R), contained more than 11% overlap with low recombination regions, which respectively accounted for 2,074 (90.17%) and 2,701 (96.22%) of the low recombination candidate windows (supplementary table S12). Notably, 97.08% (PP) and 89.17% (Pop) of candidate windows overlapping with low recombination regions were identified on chromosome XXI, which contained a large proportion (46.57%, PP; 32.16%, Pop) of the candidate windows. Recombination rates in candidate windows ($2.275 \pm 5.214 \times 10^{-3}$) were higher than the low recombination outliers ($0.380 \pm 1.245 \times 10^{-2}$) but lower than the genomic baseline ($3.369 \pm 8.415 \times 10^{-3}$) generated through a circular permutation analysis for both levels of analysis (LMM; PP, $F_{1766691} = 33000$, $P < 0.001$; Pop, $F_{586808} = 63270$, $P < 0.001$). Hence, the majority of signals of balancing selection that may arise from AOD are confined to a single chromosome in two population-pairs and three populations. Such a result suggests the majority of signals in most populations are not likely to be confounded by AOD.

In addition, we identified 13.89% (813) of candidate windows and 20.08% (195 total; 181 unique) of candidate genes at the population-pair level overlapped with F_{ST} outliers, suggesting that only a moderate proportion of signals of balancing selection could stem from population structure alone. F_{ST} was higher in candidate windows compared to both F_{ST} -outlier windows ($0.134 \pm 5.580 \times 10^{-3}$; estimate $\pm$ SE) and a genomic baseline generated through a permutation analysis ($0.451 \pm 5.557 \times 10^{-3}$, LMM; PP, $F_{956469} = 449.80$, $P < 0.001$, fig. 5C). Moreover, only 26.45% of candidate windows overlapping highly differentiated regions were located on the three chromosomes (IV, VII, and XXI) that contained the majority of signals of balancing selection (66.43% PP, 60.04% Pop). These results indicate that balancing selection can maintain high polymorphism and generate islands of divergence independently of divergent selection.

To determine whether there are functional differences between candidate genes in high and low F_{ST} or recombination regions, we ran additional GO enrichment analyses on the subsets of candidate genes overlapping and not overlapping with F_{ST} outlier or low recombination regions. Novel gene functions were enriched in the F_{ST} outlier analysis, and no new terms

identified in the recombination analysis (supplementary text S1, supplementary tables S13-S14).

Effects of Neutral Evolution and Demography

In addition to testing the performance of the genomic scans, we investigated the possibility for gene flow to maintain polymorphism by simulating five connected population-pairs under neutral evolution. We parameterised the simulation to reflect the characteristics of the empirical populations (mutation, migration, and recombination rates, and effective population sizes;^{59,65}). The simulated genome was 28.19Mb, the same size as chromosome I of the *G. aculeatus* genome. After 5,000 generations, we scanned the simulated genome for selection using T_D and θ_w in the same structured manner. We repeated this process 1,000 times for each simulated population-pair. Overall, regardless of the level of gene flow among simulated population-pairs, no candidate windows were detected in either level of analysis (supplementary table S15). In comparison, empirical chromosome I averages for T_D ranged from -0.1738 ± 0.0049 (mean $\pm$ SE) to 0.4664 ± 0.0065 , θ_w averages ranged from $1.157 \times 10^{-3} \pm 5.105 \times 10^{-6}$ to $3.122 \times 10^{-3} \pm 1.141 \times 10^{-5}$, and between 2 and 60 candidate windows were detected at the population level (supplementary table S4 and S15).

At the population-pair level, the more F_{ST} differentiated pairs generally showed more candidate windows than less differentiated population-pairs (table 1; fig 3A-B). Hence, to determine the number of simulated F_{ST} outliers that emerge from population structure alone, we identified outliers in the simulated data using thresholds determined from the top 10% of empirical data (CA, $F_{ST} > 0.570$; G1, $F_{ST} > 0.405$; G2, $F_{ST} > 0.272$; NO, $F_{ST} > 0.380$; US, $F_{ST} > 0.218$). Mean numbers of outlier windows ranged from 0.65 (CA) to 213.19 (US; supplementary table S15). Higher numbers of F_{ST} outliers were detected in simulated population-pairs with lower migration due to lower empirical thresholds being used. In contrast, we detected between 1,916 (CA) to 2,580 (NO) F_{ST} outliers on chromosome I in the empirical genome scan (supplementary table S2). Together, these results highlight the empirical signals we observe cannot solely be explained by neutral evolution or population structure in connected population-pairs, and therefore stem from natural selection or other demographic effects.

We investigated whether patterns of rapid demography change could impact the detection of balancing selection by performing coalescent simulations using a Multiple Sequential Markovian Coalescent approach (MSMC;⁶¹). Among the 10 populations, changes in historical effective population sizes (N_e) fall into two broad categories. The Norwegian lake population (NO_L) showed stable or gradual increases in N_e , while the nine other populations showed gradual decrease in N_e since the last glaciations (ca. 11 kya, fig. 6). These decreases could alter allele frequencies mimicking signals of balancing selection, albeit to a lesser extent than

if bottlenecks had occurred. Hence, we cannot exclude that demography affected signals of balancing selection in our data, but the possible effects are likely minor.

Discussion

Balancing selection maintains genetic diversity and fuels the evolutionary potential of species and populations^{4,66}. In this study, we present a comprehensive analysis of genome-wide signals of balancing selection in *G. aculeatus*, accounting for the impact of missing data, population structure, associated overdominance, gene flow and demography. We demonstrate that balancing selection is frequent, with candidate regions characterised by an enrichment of nonsynonymous SNPs. We also observed LD to be significantly extended around putative sites of selection, especially along multiple contiguous candidate windows and population-specific candidate genes, suggesting the majority of balancing selection in *G. aculeatus* is recent. The functions of candidate genes and their often population-specific nature further suggest balancing selection may play an important role in adaptation to heterogeneous environments. Finally, by utilising diverse metrics, placing results in a demographic framework, and simulating neutral evolution, we considered potential confounding factors like other forms of natural selection, migration, population structure, and associated overdominance (AOD).

Targets of Balancing Selection in *G. aculeatus*

Despite long-standing interest in balancing selection, the extent of genes affected and their putative functions remain elusive^{4,9}. We identified between 715 (Population-pair) and 1,010 (Population) unique protein-coding genes with signals of balancing selection, comprising 3.44%-4.86% of all protein-coding genes, spanning between 2.31%-3.10% of the genome. Comparatively, other scans for balancing selection identified: 78 (1.5% of coding genes) genes in the plant pathogen *R. solanacearum*³⁴; 99 (3.8%) genes in the *S. aureus* genome³⁵; 183 (1.3%) genes across the *D. melanogaster* genome²⁴; and between 1,404-1,616 (6.7%-7.8%) genes in humans³³. Notably, genome-wide signals of balancing selection across studies tend to increase when more populations are included in the analysis, suggesting the majority of candidate genes would be found at the population level. Interestingly, we detected population-specific enrichment of SNP functions, whereby multiple populations showed enrichment of nonsynonymous and intergenic SNPs and one population contained enrichment of SNPs in regulatory elements. Such a result is compatible with the scenario that balancing selection acts primarily on protein-coding sites, but also can affect gene expression by accumulating functional variants in regulatory regions.

Non-Random Distribution of Candidate Genes

Recent work investigating the interplay between ecological factors, morphological traits, and genomic differentiation, has revealed that the loci that underlie adaptation in *G. aculeatus* are primarily located on a few chromosomes^{53,67}. Here, the majority of candidate windows and genes were identified on chromosomes IV, VII, and XXI. In the three-spined stickleback, chromosomes IV and XXI are enriched for quantitative trait loci (QTL) of diverse types (feeding, shape, defence, behaviour, swimming, pigmentation, respiration, body size, and reproduction), whereas chromosome VII is enriched in defence trait QTLs (dorsal spine and lateral plate morphology;^{53,67,68}). Chromosomes IV and VII are also among chromosomes most enriched with deletions and duplications⁵⁸, and novel genes⁶⁹. It should be noted that genetic diversity on chromosome XXI may equally likely have arisen from AOD and balancing selection, whereas only 26.45% of candidate windows that may be confounded by population structure were located on chromosomes IV, VII, and XXI. Hence, the abundance of signals which can only reasonably be explained by balancing selection on chromosomes IV and VII suggests our scan captured signals of adaptive loci evolving under balancing selection.

Population-Specificity of Balancing Selection

The parapatric population pairs have enabled us to compare the contribution of balancing selection at different levels (i.e., population and population-pair). We show that the majority of candidate genes were identified at the population level (59.55%). This pattern suggests the local ecological factors unique to each population maintains genetic polymorphism. Overall, we found an enrichment for genes involved in sensory perception of smell as well as glycosaminoglycan (GAG) catalytic process genes. In particular, GAGs are involved in cell signalling, growth, adhesion, migration, and differentiation, and are found ubiquitously on the cell surface, in intracellular compartments, and in the extracellular matrix⁷⁰. GAGs expressed on cell surfaces are implicated in pathogenesis and cell invasion in microbial infection⁷⁰, and have been observed as potential targets of balancing selection previously⁷¹.

Population-specific patterns of selection are becoming more prominent in the literature, particularly so in *G. aculeatus* for genomic islands of differentiation⁵⁹, CNVs and novel genes^{58,69}, as well as genome-wide genotype-expression relationships⁷². Because polymorphism fuels adaptation, it is regularly assumed that population-specific genetic diversity is associated with local adaptation. While the adaptive nature of this genetic diversity would need investigating (e.g.,⁴⁸), we have revealed the level at which balancing selection operates the strongest.

Instances of balancing selection across the same loci in multiple populations suggest the presence of stable long-term or recurrent parallel selective pressures and the adaptive nature of the response^{23,33,73}. We identified 211 (PP) and 272 (Pop) parallel candidate genes across

both levels of analysis. The highest proportion of parallel candidate genes at the population level were shared among populations belonging to the same ecotype (i.e., river or lake stickleback; 38.36%). This is particularly notable considering the relatively low proportion (4.55%) of candidate genes shared among geographically connected population-pairs. Those candidate genes identified multiple times serve as strongest evidence of both the action of balancing selection and its adaptive nature.

At the population level, parallel candidate genes and the subset known to be copy-number polymorphic⁵⁸ showed functional enrichment of olfactory genes. The nine candidate olfactory genes from four chromosomes (I, V, VII, and XVI) are all orthologous to *D. rerio* Olfactory Receptor (OR) superfamily genes. Since each olfactory neuron only expresses a single allele from a single OR gene, higher levels of diversity in OR genes should correspondingly increase olfactory acuity⁷⁴. This increased acuity may enhance the many roles olfactory discrimination plays in aquatic systems: conspecific recognition and grouping behaviour⁷⁵, homing and spawning behaviours⁷⁶, olfactory-based mate choice^{77–79} and foraging and predator avoidance^{80,81}. Importantly, genetic diversity in CNV regions is likely the result of mechanisms other than balancing selection⁸².

We identified *Plxnc1* as a candidate gene in all European populations studied herein. The exact role of *Plxnc1* in olfaction remains unclear in fish, but recent evidence demonstrates it triggers post-synaptic assembly in the mouse olfactory system⁸³. Interestingly, *G. aculeatus* exhibits MHC-based olfactory mate choice where females choose males both to optimise MHC diversity and based on a currently unidentified “male validation factor”^{84,85}. Here we speculate that *Plxnc1* could be co-opted to scrutinise *G. aculeatus* sexual signals, further linking the maintenance of polymorphism and sexual selection.

Characteristics of Candidate Windows

Theory predicts that balancing selection should extend linkage disequilibrium (LD) around targets of recent selection^{4,38,86,87}. Here, we observe extended LD up to 150kb away from loci putatively under balancing selection. We also showed LD is most extended around wider contiguous candidate windows. Stronger LD near the site of selection in all populations indicates balancing selection in *G. aculeatus* is predominantly recent. As such, heterozygote advantage/overdominance and negative-frequency dependent selection are likely essential in maintaining LD faster than recombination can remove it. Interestingly, private candidate genes, those most associated with population-specific selection, had more extended LD than more stable and putatively older parallel candidate genes. Observations of overdominance or negative-frequency dependent selection have been made for olfactory genes⁸⁸, and more specifically within *G. aculeatus* immune genes⁵ confirming these types of selection operate at the investigated scales.

Neutrality tests can be confounded by multiple factors that need to be accounted for⁹. One such factor is AOD, which can generate genetic diversity in regions of low recombination around linked loci therein mimicking signals of balancing selection^{40,41}. Alongside extended LD, candidate windows and genes were only associated with a small overlap with conservatively estimated low recombination regions in 3 of 5 population-pairs and 8 of 10 populations. In the remaining population-pairs and populations, however, a high proportion of candidate windows overlapped with low recombination regions and may also arise from AOD as well as balancing selection. Interestingly, the majority of signals of AOD were observed on chromosome XXI in three river populations. In general, we also detected more signals of balancing selection using neutrality statistics in these populations and on chromosome XXI, suggesting this genetic diversity may be maintained by mechanisms other than balancing selection. It should be noted that AOD acts in a similar manner to true overdominance⁴⁰, and thus the overlap with regions of low recombination is only suggestive that AOD may be operating in those regions.

Another factor that may confound signals of natural selection is population structure^{9,64}. And whilst candidate windows were characterised by slightly elevated F_{ST} in comparison to the genomic baseline, there was only a small overlap between conservative estimates of F_{ST} outliers and candidate windows and genes. Such results suggest the balancing selection signals detected in the majority of the populations and pairs do not stem from population structure and divergent selection. Based on these observations, we suggest balancing selection can form the basis of islands of divergence, maintaining standing genetic variation upon which natural selection can act^{7,77}. Furthermore this work highlights how the origins of differentiation may arise through a mosaic of mechanisms, including balancing selection, other forms of natural selection^{53,68}, soft sweeps and AOD in addition of demographic changes occurring at different time scales.

Effects of Demography

It is well described that demography impacts signals of balancing selection. Scrutinising the proportion of candidate windows and genes overlapping with F_{ST} outliers, our simulations show that spurious signals from migration alone are rare. Similarly, using coalescent simulations, we observed gradual N_e declines in almost all populations (with the exception of NO_L). Consequently, for these populations, demography may have resulted in slight increased polymorphism at some loci not under balancing selection. Our results highlight the importance of investigating population demography, both in terms of sampling (e.g., including multiple ecologically distinct populations and scanning across multiple levels of organization) and downstream analyses (e.g., demographic simulations and utilising multiple lines of evidence).

Conclusions

Overall, our results show many elements of balancing selection are population-specific likely emerging from heterogeneous local habitats. This population-specific polymorphism from the evolutionary potential of species and fuels adaptive evolution. We showed that signal of balancing selection is neither rare nor randomly distributed across the genome of the three-spined stickleback.

Methods

Sampling and Data Processing

We used 60 complete resequenced genomes of *G. aculeatus* from 5 recently diverged freshwater lake and river pairs (table S1; ERP004574; details about sampling, library preparation, and sequencing can be found in a previous study⁵⁹. The five population pairs were connected lake-river systems in Canada, the United States, Norway, and two sites in Germany. A mixture of small-insert, large-insert, and mate-pair libraries (100bp reads with 140bp and 300bp insert sizes, and 50bp reads with 3kb insert size, respectively) were sequenced for each individual^{58,59}.

Adapter cleaned reads were trimmed using Trimmomatic v0.36⁸⁹ in paired-end mode, trimming read tails with a PHRED quality score below 20, trimming to a maximum of half the read. Reads were mapped to the *G. aculeatus* reference genome (Broad/gasAcu1) with BWA-MEM v0.7.17⁹⁰. Mate-pair information was verified, and reads were sorted and merged using the Picard Toolkit (<http://broadinstitute.github.io/picard/>). A first variant call was conducted using GATK v4.0⁹¹ following the *best practices* workflow⁹². Because information on variant and non-variant sites is required for the MSMC analysis⁶¹, a second variant call was performed using GATK v3.8 to additionally include non-variant sites in the VCF file. Both calls used a multi-sample approach, whereby all genomes were called simultaneously, split by chromosome. The final SNP sets used for the genome scan was produced by filtering variants with the following thresholds “QD < 5.0 | MQ < 50 | FS > 60 | SOR > 3.0 | QUAL < 30”. The MSMC SNP set additionally filtered each sample individually based on the distribution of coverage for each sample, removing the bottom 5% and top 10% of coverage distribution, and removing all variants within 10bp of an indel called in GATK. The MSMC SNP set genotypes were phased using the read-aware approach in SHAPEIT2⁹³. Because MSMC can only analyse biallelic sites, all multiallelic sites were also removed from the MSMC SNP set, leaving a total of 7,761,294 SNPs. Genome mapping and variant calls were conducted on the QMUL Apocrita High Performance Computing Cluster⁹⁴.

Primary Genome Scans

Genome scans were analysed in R v3.6.1⁹⁵, and population genetic indices were calculated using the R package 'PopGenome'⁹⁶. We used Tajima's D ⁴² and Watterson's estimator (θ_w)⁶⁰ to identify signals of balancing selection, and π and NCD1³³ to verify signals of selection. We calculated all metrics in overlapping 7.5kb sliding windows with step size 1.5kb across the genome. Because diversity metrics are influenced by missing information, we masked genomic windows with over 1% missing information, including missing genotypes and no coverage at segregating sites for a sample. We performed this analysis across two levels of organisation, grouping individuals by population-pair and by population. F_{ST} was additionally calculated within genomic windows only at the population-pair level. We chose relative differentiation (F_{ST}) over absolute divergence (D_{XY}) due to the latter being unreliable for nascent population-pairs in non-equilibrium during differentiation. Here, we define NCD1 as the deviation from the target allele frequency of 0.5 ($t \neq 0.5$). Various window sizes and overlapping step sizes were tested and a combination was chosen as a compromise to accurately identify the origin of the selection signal (supplementary fig. S6). For the genome scan, we used all 20 autosomal chromosomes and the unmapped chromosome, which contains all unmapped scaffolds separated by 1000 N's. When multiple contiguous windows met the candidate window threshold, diversity metrics were averaged across the windows, and downstream analyses only used the start and end coordinates of the contiguous windows to reduce statistical autocorrelation.

Functional SNP Annotations

Functional annotations were obtained for SNPs using the Ensembl Variant Effect Predictor from Ensembl version 95⁹⁷. SNPs were categorised as intergenic, exonic, intronic, regulatory, and splice site, and the sum of SNPs in each category was calculated across all candidate windows. Additionally, we partitioned exonic SNPs into either synonymous or nonsynonymous categories to evaluate whether protein structure was targeted by selection. A null distribution was generated for each population-pair and population by using a circular genomic permutation approach⁹⁸. Briefly, the genome was considered to be circular, ordered from chromosome I to chromosome XXI, and restarting again at chromosome I. For each permutation, the observed functional annotations were permuted by rotation with respect to their genomic locations by a random number between 1 and the number of SNPs. This approach retains the original order and structure but uncouples the association between SNPs and functional annotation to generate a null distribution. A total of 1,000 permutations was used to generate the null distribution. We obtained empirical one-tailed P -values by comparing the distribution of candidate SNP effects to the null distributions for each population-pair and population.

GO Enrichment Analysis

We used the Ensembl version 95 *G. aculeatus* gene annotations to obtain candidate gene orthologs between *G. aculeatus* and *Danio rerio*. A total of 16,027 *G. aculeatus* coding genes (77.10%) were matched to 14,762 unique zebrafish orthologs, forming the basis for Gene Ontology (GO) enrichment analyses. Orthologs were used because *D. rerio* has the most extensive and best functional annotation of any fish species.

GO enrichment for candidate genes was calculated in R using the topGO package⁹⁹ from the Bioconductor project¹⁰⁰. Only GO terms supported by a minimum of 5 annotated genes were included in the GO enrichment analysis⁹⁹. For both levels of organisation, enrichment was estimated using a universe of all autosomal genes, significance was inferred using a weighted Fisher's exact test, and *P*-values were corrected for multiple testing using FDR. A GO enrichment analysis was also performed on candidate copy number variable genes identified at the population level using CNV regions previously documented in these genomes⁵⁸. In addition, to identify any functional biases in GOs in the *G. aculeatus* genome, a null distribution of was generated using circular genomic permutations. The observations for each of the 1,000 permutations was the number and genomic structure of candidate genes identified at the population level. This analysis was repeated at the population-pair level. Empirical *P*-values for enriched functions were obtained through one-tailed tests of the proportion of enriched functions identified more than expected by chance ($P > 0.05$). "Sensory Perception of Smell" was the most common term (excluding "Biological Process") with a one-tailed *P*-value of 0.019 and 0.020 across the population-pair and population levels, respectively (supplementary table S16).

Performance of Genome Scans

The performance of the genome scan was analysed using linear mixed-effect models in the lme4 package¹⁰¹, and *P*-values were calculated using Satterthwaite's degree of freedom method implemented in the lmerTest package¹⁰². Firstly, a null distribution of π and NCD1 was estimated by randomly subsampling 1000 genomic windows for every candidate window from the same population and chromosome. Using either π or NCD1 as a response variable, we compared candidate windows with the null distribution using population as a random effect in a mixed-effect model.

Linkage Disequilibrium, Associated Overdominance, and Differentiation

Estimates of the squared correlation coefficient (R^2) between pairs of SNPs were obtained using PLINK v1.90¹⁰³ for each population separately. Two SNP matrices were constructed for each chromosome and population: (i) genomic LD background, and (ii) LD around candidate windows. The first matrix consisted of all SNPs present, and the second matrix was used to measure linkage only around SNPs within candidate windows. We ran PLINK with the

parameters “--ld-window-r2 0 --ld-window-kb 3000 --ld-window 1000”, enabling all R^2 to be reported below the default threshold. Next, we assigned R^2 values to 500bp bins according to physical distances between SNP pairs. To visualise differences in LD decay between the genomic baseline and genomic regions surrounding sites under putative balancing selection, we calculated ΔR^2 as the mean R^2 value of candidate windows minus the mean R^2 value derived from genomic regions surrounding candidate windows. We then used mean R^2 for all pairwise bins up to 200kb as a response variable, a fixed effect whether the sample was from a lake or river population, and population as a random effect. To identify when extended LD was no longer detectable, we grouped mean R^2 for all bins within 0kb-10kb as a response variable, a fixed effect whether the R^2 values were from the genomic baseline or surrounding sites under balancing selection, and defined population as a random effect. We repeated this analysis for all non-overlapping 10kb windows up to 190kb-200kb and P -values were corrected using FDR. Finally, we investigated patterns in LD around private candidate genes, parallel candidate genes, quasi-candidate genes (non-candidate genes that are candidates in one or more other populations), and the genomic baselines. For this analysis we randomly subsampled the 949 genes (number of candidate genes across the autosomes, excluding the unmapped chromosome) to generate the genomic LD null distribution, then calculated LD around all 949 candidate genes in every population, and separated results into candidate genes, parallel candidate genes, and quasi-candidate genes.

To account for potentially confounding signals arising from population structure or AOD rather than balancing selection, a baseline distribution of F_{ST} and recombination rates were generated using a circular genomic permutation approach. Recombination rate was averaged across intervals defined by adjacent markers in the genetic map⁶⁵, which was assigned to overlapping with candidate windows. The number of observations for each of the 1,000 permutations in each of the analyses matched the number of candidate windows either overlapping with F_{ST} outliers for each population-pair, or candidate windows in each population-pair or population for the AOD analyses. The permutation approach was repeated for each population-pair and population individually. Differences between the observed F_{ST} and recombination rate distributions in candidate windows were compared to the null distribution using mixed effects models, using population-pairs or population as the random effect.

Neutral Evolution and Demography

To test whether the signals we detected stem from balancing selection, neutral evolution, or gene flow, we simulated all five population-pairs evolving under neutral evolution using SLiM v3.3¹⁰⁴. The initial Wright-Fisher populations were parametrised to reflect the real population-pairs where possible, using a genome-wide recombination rate of 3.11 cM/Mb⁶⁵, a mutation

rate of $1.4812 \times 10^{-08;59}$, initial effective population sizes were based on previous estimates⁵⁹, and the near-symmetric migration rate was calculated from the genome-wide F_{ST} values. The simulated genome was the same size as the *G. aculeatus* chromosome I (28.2Mb). Neutral evolution was simulated for 5,000 generations, after which 6 individuals were randomly sampled from each population. A custom R script was used to process the mutations and apply them to *G. aculeatus* chromosome I, which was used to scan for signals of selection in the same structured manner as above, i.e. across population-pairs and populations. We repeated this analysis 1,000 times for each simulated population-pair.

In addition to estimating the impact of gene flow using simulations, we assessed the possible impact of historic population-size fluctuations using MSMC⁶¹, which estimates demographic histories. Using four individuals randomly chosen from each population (supplementary table S1), variants were processed into the MSMC format using the *vcfAllSiteParser* and *generatemultihetsep* scripts from the msmc-tools repository (<https://github.com/stschiff/msmc-tools>). Because MSMC relies on accurate detection of heterozygous sites, only samples with an average genome-side coverage >15x were retained. Additionally, a mappability mask was generated with SNPable (<http://lh3lh3.users.sourceforge.net/snpable.shtml>; k=100 and r=0.5), masking regions where the overlapping 100-mers were not mapped uniquely. We ran MSMC using a mutation rate of 1.4812×10^{-8} per site per generation⁵⁹, a generation time of 1 year, a fixed recombination of 3.11 cM/Mb (Roesti et al. 2013; setting “rhoovermu = 2”), and otherwise using the default parameters. Only the 20 autosomal chromosomes were included in this analysis.

Data Availability

The genome sequences of the 60 *G. aculeatus* fish were obtained from previous publications^{58,59}, and retrieved from the European Nucleotide Archive with the accession number ERP004574.

Scripts used to conduct the genome scan and statistical analyses are available online (https://github.com/dthorburn/Balancing_Selection_Scan).

Acknowledgements

We thank Queen Mary University of London ITS Research support for technical assistance, and Richard Nichols for discussions about the neutral evolution simulations. We would also like to thank the editor and two anonymous reviewers for their constructive and insightful comments on the previous version of this manuscript.

Authors' contributions

C.E. and D.M.J.T. conceived and designed the study. F.J.J.C., P.G.D.F., C.E., T.L.L. obtained the samples and oversaw the sequencing. T.B.H.R., E.B.B. and M.M. created the Big Screen

Consortium with grant support from the Max Planck Society to M.M. D.M.J.T. analysed the data, with technical support from K.S. C.E. advised the analyses. D.M.J.T. and C.E. drafted the manuscript. All co-authors contributed to the final version of the manuscript. C.E. acknowledges the support of the German Science Foundation (DFG, EI841/4-1 and EI 841/6-1). D.M.J.T. was supported by a studentship from Queen Mary University of London.

References

1. Dobzhansky, T. A review of some fundamental concepts and problems of population genetics. *Cold Spring Harb. Symp. Quant. Biol.* **20**, 1–15 (1955).
2. Kimura, M. Evolutionary rate at the molecular level. *Nature* **217**, 624–626 (1968).
3. Nei, M. Selectionism and neutralism in molecular evolution. *Mol. Biol. Evol.* **22**, 2318–2342 (2005).
4. Charlesworth, D. Balancing selection and its effects on sequences in nearby genome regions. *PLoS Genet.* **2**, 379–384 (2006).
5. Eizaguirre, C., Lenz, T. L., Kalbe, M. & Milinski, M. Rapid and adaptive evolution of MHC genes under parasite selection in experimental vertebrate populations. *Nat. Commun.* **3**, 621–626 (2012).
6. Barrett, R. D. H. & Schluter, D. Adaptation from standing genetic variation. *Trends Ecol. Evol.* **23**, 38–44 (2008).
7. Marques, D. A., Meier, J. I. & Seehausen, O. A Combinatorial View on Speciation and Adaptive Radiation. *Trends Ecol. Evol.* **34**, 531–544 (2019).
8. Spurgin, L. G. & Richardson, D. S. How pathogens drive genetic diversity: MHC, mechanisms and misunderstandings. *Proc. R. Soc. B Biol. Sci.* **277**, 979–988 (2010).
9. Fijarczyk, A. & Babik, W. Detecting balancing selection in genomes: Limits and prospects. *Mol. Ecol.* **24**, 3529–3545 (2015).
10. Cox, R. M. & Calsbeek, R. Sexually antagonistic selection, sexual dimorphism, and the resolution of intralocus sexual conflict. *Am. Nat.* **173**, 176–187 (2009).
11. Levene, H. Genetic Equilibrium When More Than One Ecological Niche is Available. *Am. Nat.* **87**, 331–333 (1953).
12. Charbonnel, N. & Pemberton, J. A long-term genetic survey of an ungulate population reveals balancing selection acting on MHC through spatial and temporal fluctuations in selection. *Heredity (Edinb)*. **95**, 377–388 (2005).

- 680 13. Hedrick, P. W. Pathogen resistance and genetic variation at MHC loci. *Evolution* (N.
681 Y). **56**, 1902–1908 (2002).
- 682 14. Wheat, C. W., Haag, C. R., Marden, J. H., Hanski, I. & Frilander, M. J. Nucleotide
683 polymorphism at a gene (Pgi) under balancing selection in a butterfly metapopulation.
684 *Mol. Biol. Evol.* **27**, 267–281 (2010).
- 685 15. Mérot, C., Llaurens, V., Normandeau, E., Bernatchez, L. & Wellenreuther, M.
686 Balancing selection via life-history trade-offs maintains an inversion polymorphism in
687 a seaweed fly. *Nat. Commun.* **11**, 1–11 (2020).
- 688 16. Foerster, K. *et al.* Sexually antagonistic genetic variation for fitness in red deer.
689 *Nature* **447**, 1107–1110 (2007).
- 690 17. Lundberg, M., Zhong, X., Konrad, A., Olsen, R. A. & Råberg, L. Balancing selection in
691 Pattern Recognition Receptor signalling pathways is associated with gene function
692 and pleiotropy in a wild rodent. *Mol. Ecol.* **29**, 1990–2003 (2020).
- 693 18. Wegner, K. M., Reusch, T. B. H. & Kalbe, M. Multiple parasites are driving major
694 histocompatibility complex polymorphism in the wild. *J. Evol. Biol.* **16**, 224–232
695 (2003).
- 696 19. Stahl, E. A., Dwyer, G., Mauricio, R., Kreitman, M. & Bergelson, J. Dynamics of
697 disease resistance polymorphism at the Rpm1 locus of *Arabidopsis*. *Nature* **400**, 667–
698 671 (1999).
- 699 20. Penn, D. J., Damjanovich, K. & Potts, W. K. MHC heterozygosity confers a selective
700 advantage against multiple-strain infections. *Proc. Natl. Acad. Sci. U. S. A.* **99**,
701 11260–11264 (2002).
- 702 21. Takahata, N. & Nei, M. Allelic genealogy under overdominant and frequency-
703 dependent selection and polymorphism of major histocompatibility complex loci.
704 *Genetics* **124**, 967–987 (1990).
- 705 22. Arora, J. *et al.* HLA Heterozygote Advantage against HIV-1 Is Driven by Quantitative
706 and Qualitative Differences in HLA Allele-Specific Peptide Presentation. *Mol. Biol.*
707 *Evol.* **37**, 639–650 (2020).
- 708 23. Unckless, R. L., Howick, V. M. & Lazzaro, B. P. Convergent Balancing Selection on
709 an Antimicrobial Peptide in *Drosophila*. *Curr. Biol.* **26**, 257–262 (2016).
- 710 24. Croze, M. *et al.* A genome-wide scan for genes under balancing selection in
711 *Drosophila melanogaster*. *BMC Evol. Biol.* **17**, 1–12 (2017).

- 712 25. Wright, S. The distribution of self-sterility alleles in populations. *Genetics* **4**, 538–552
713 (1939).
- 714 26. Roux, C. *et al.* Recent and Ancient Signature of Balancing Selection around the S-
715 Locus in *Arabidopsis halleri* and *A. lyrata*. **30**, 435–447 (2012).
- 716 27. Cho, S., Huang, Z. Y., Green, D. R., Smith, D. R. & Zhang, J. Evolution of the
717 complementary sex-determination gene of honey bees: Balancing selection and trans-
718 species polymorphisms. *Genome Res.* **16**, 1366–1375 (2006).
- 719 28. Ségurel, L. *et al.* The ABO blood group is a trans-species polymorphism in primates.
720 *Proc. Natl. Acad. Sci. U. S. A.* **109**, 18493–18498 (2012).
- 721 29. Barson, N. J. *et al.* Sex-dependent dominance at a single locus maintains variation in
722 age at maturity in salmon. *Nature* **528**, 405–408 (2015).
- 723 30. Sabeti, P. C. *et al.* Genome-wide detection and characterization of positive selection
724 in human populations. *Nature* **449**, 913–918 (2007).
- 725 31. Mu, J. *et al.* *Plasmodium falciparum* genome-wide scans for positive selection,
726 recombination hot spots and resistance to antimalarial drugs. *Nat. Genet.* **42**, 268–
727 271 (2010).
- 728 32. Oleksyk, T. K., Smith, M. W. & O’Brien, S. J. Genome-wide scans for footprints of
729 natural selection. *Philos. Trans. R. Soc. B Biol. Sci.* **365**, 185–205 (2010).
- 730 33. Bitarello, B. D. *et al.* Signatures of long-term balancing selection in human genomes.
731 *Genome Biol. Evol.* **10**, 939–955 (2018).
- 732 34. Castillo, J. A. & Agathos, S. N. A genome-wide scan for genes under balancing
733 selection in the plant pathogen *Ralstonia solanacearum*. *BMC Evol. Biol.* **19**, 1–16
734 (2019).
- 735 35. Thomas, J. C., Godfrey, P. A., Feldgarden, M. & Robinson, D. A. Candidate targets of
736 balancing selection in the genome of *Staphylococcus aureus*. *Mol. Biol. Evol.* **29**,
737 1175–1186 (2012).
- 738 36. Leffler, E. M. *et al.* Multiple Instances of Ancient Balancing Selection Shared Between
739 Humans and Chimpanzees. *Science (80-.).* **339**, 1578–1582 (2013).
- 740 37. Asthana, S., Schmidt, S. & Sunyaev, S. A limited role for balancing selection. *Trends*
741 *Genet.* **21**, 30–32 (2005).
- 742 38. Andrés, A. M. *et al.* Targets of balancing selection in the human genome. *Mol. Biol.*

743 *Evol.* **26**, 2755–2764 (2009).

744 39. Gao, Z., Przeworski, M. & Sella, G. Footprints of ancient-balanced polymorphisms in
745 genetic variation data from closely related species. *Evolution (N. Y.)*. **69**, 431–446
746 (2015).

747 40. Gilbert, K. J., Pouyet, F., Excoffier, L. & Peischl, S. Transition from Background
748 Selection to Associative Overdominance Promotes Diversity in Regions of Low
749 Recombination. *Curr. Biol.* **30**, 101–107.e3 (2020).

750 41. Becher, H., Jackson, B. C. & Charlesworth, B. Patterns of Genetic Variability in
751 Genomic Regions with Low Rates of Recombination. *Curr. Biol.* **30**, 94–100.e3 (2020).

752 42. Tajima, F. Statistical method for testing the neutral mutation hypothesis by DNA
753 polymorphism. *Genetics* **123**, 585–595 (1989).

754 43. McDonald, J. H. & Kreitman, M. Adaptive protein evolution at the *Adh* locus in
755 *Drosophila*. *Nature* **351**, 652–654 (1991).

756 44. Zhen, Y. & Andolfatto, P. Methods to detect selection on noncoding DNA. *Methods*
757 *Mol. Biol.* **856**, 141–159 (2012).

758 45. Feulner, P. G. D. *et al.* Genome-wide patterns of standing genetic variation in a
759 marine population of three-spined sticklebacks. *Mol. Ecol.* **22**, 635–649 (2013).

760 46. Durmaz, E. *et al.* A clinal polymorphism in the insulin signaling transcription factor
761 *foxo* contributes to life-history adaptation in *Drosophila*. *Evolution (N. Y.)*. **73**, 1774–
762 1792 (2019).

763 47. Bellis, E. S. *et al.* Genomics of sorghum local adaptation to a parasitic plant. *Proc.*
764 *Natl. Acad. Sci. U. S. A.* **117**, 4243–4251 (2020).

765 48. Eizaguirre, C., Lenz, T. L., Kalbe, M. & Milinski, M. Divergent selection on locally
766 adapted major histocompatibility complex immune genes experimentally proven in the
767 field. *Ecol. Lett.* **15**, 723–731 (2012).

768 49. Fang, B., Merilä, J., Ribeiro, F., Alexandre, C. M. & Momigliano, P. Molecular
769 Phylogenetics and Evolution Worldwide phylogeny of three-spined sticklebacks. *Mol.*
770 *Phylogenet. Evol.* **127**, 613–625 (2018).

771 50. McKinnon, J. S. & Rundle, H. D. Speciation in nature: The threespine stickleback
772 model systems. *Trends Ecol. Evol.* **17**, 480–488 (2002).

773 51. Bell, M. A. & Foster, S. A. The evolutionary biology of the threespine stickleback.

- 774 *Oxford Univ. Press* **584**, (1994).
- 775 52. Roesti, M., Kueng, B., Moser, D. & Berner, D. The genomics of ecological vicariance
776 in threespine stickleback fish. *Nat. Commun.* **6**, 1–14 (2015).
- 777 53. Rennison, D., Stuart, Y., Bolnick, D. & Peichel, C. Ecological factors and
778 morphological traits are associated with repeated genomic differentiation between
779 lake and stream stickleback. *Philos. Trans. R. Soc. B Biol. Sci.* **374**, 1–8 (2019).
- 780 54. Stuart, Y. E. *et al.* Contrasting effects of environment and genetics generate a
781 continuum of parallel evolution. *Nat. Ecol. Evol.* **1**, 1–7 (2017).
- 782 55. Eizaguirre, C. *et al.* Parasite diversity, patterns of MHC II variation and olfactory
783 based mate choice in diverging three-spined stickleback ecotypes. *Evol. Ecol.* **25**,
784 605–622 (2011).
- 785 56. Huang, Y. *et al.* Transcriptome profiling of immune tissues reveals habitat-specific
786 gene expression between lake and river sticklebacks. *Mol. Ecol.* **25**, 943–958 (2016).
- 787 57. Ravinet, M., Prodöhl, P. A. & Harrod, C. Parallel and nonparallel ecological,
788 morphological and genetic divergence in lake-stream stickleback from a single
789 catchment. *J. Evol. Biol.* **26**, 186–204 (2013).
- 790 58. Chain, F. J. J. *et al.* Extensive Copy-Number Variation of Young Genes across
791 Stickleback Populations. *PLoS Genet.* **10**, 1–18 (2014).
- 792 59. Feulner, P. G. D. *et al.* Genomics of Divergence along a Continuum of Parapatric
793 Population Differentiation. *PLoS Genet.* **11**, 1–18 (2015).
- 794 60. Watterson, G. A. On the number of segregating sites in genetical models without
795 recombination. *Theor. Popul. Biol.* **7**, 256–276 (1975).
- 796 61. Schiffels, S. & Durbin, R. Inferring human population size and separation history from
797 multiple genome sequences. *Nat. Genet.* **46**, 919–925 (2014).
- 798 62. Jones, F. C. *et al.* The genomic basis of adaptive evolution in threespine sticklebacks.
799 *Nature* **484**, 55–61 (2012).
- 800 63. Giner-Delgado, C. *et al.* Evolutionary and functional impact of common polymorphic
801 inversions in the human genome. *Nat. Commun.* **10**, 1–14 (2019).
- 802 64. Nielsen, R. Molecular Signatures of Natural Selection. *Annu. Rev. Genet.* **39**, 197–
803 218 (2005).
- 804 65. Roesti, M., Moser, D. & Berner, D. Recombination in the threespine stickleback

805 genome - Patterns and consequences. *Mol. Ecol.* **22**, 3014–3027 (2013).

806 66. Eizaguirre, C. & Baltazar-soares, M. Evolutionary conservation — evaluating the
807 adaptive potential of species. *Evol. Appl.* **7**, 963–967 (2014).

808 67. Peichel, C. L. & Marques, D. A. The genetic and molecular architecture of phenotypic
809 diversity in sticklebacks. *Philos. Trans. R. Soc. B Biol. Sci.* **372**, 1–8 (2017).

810 68. Rennison, D., Rudman, S. & Schluter, D. Genetics of adaptation: Experimental test of
811 a biotic mechanism driving divergence in traits and genes. *Evol. Lett.* **3**, 513–520
812 (2019).

813 69. Schmitz, J. F., Chain, F. J. J. & Bornberg-Bauer, E. Evolution of novel genes in three-
814 spined stickleback populations. *Heredity (Edinb)*. **125**, 50–59 (2020).

815 70. Kamhi, E., Joo, E. J., Dordick, J. S. & Linhardt, R. J. Glycosaminoglycans in infectious
816 disease. *Biol. Rev.* **88**, 928–943 (2013).

817 71. Somel, M. *et al.* A scan for human-specific relaxation of negative selection reveals
818 unexpected polymorphism in proteasome genes. *Mol. Biol. Evol.* **30**, 1808–1815
819 (2013).

820 72. Huang, Y. *et al.* Genome-Wide Genotype-Expression Relationships Reveal Both
821 Copy Number and Single Nucleotide Differentiation Contribute to Differential Gene
822 Expression between Stickleback Ecotypes. *Genome Biol. Evol.* **11**, 2344–2359
823 (2019).

824 73. Sweeney, C. G. *et al.* Convergent Balancing Selection on the Mu-Opioid Receptor in
825 Primates. *Mol. Biol. Evol.* **34**, 1629–1643 (2017).

826 74. Shykind, B. M. Regulation of odorant receptors: One allele at a time. *Hum. Mol.*
827 *Genet.* **14**, 33–39 (2005).

828 75. Hiermes, M., Mehlis, M., Rick, I. P. & Bakker, T. C. M. Habitat-dependent olfactory
829 discrimination in three-spined sticklebacks (*Gasterosteus aculeatus*). *Anim. Cogn.*
830 839–846 (2015). doi:10.1007/s10071-015-0850-8

831 76. Hasler, A. D., Scholz, A. T. & Horrall, R. M. Olfactory Imprinting and Homing in
832 Salmon. *Am. Sci.* **66**, 347–355 (1978).

833 77. Eizaguirre, C., Yeates, S. E., Lenz, T. L., Kalbe, M. & Milinski, M. MHC-based mate
834 choice combines good genes and maintenance of MHC polymorphism. *Mol. Ecol.* **18**,
835 3316–3329 (2009).

- 836 78. Plenderleith, M., Van Oosterhout, C., Robinson, R. L. & Turner, G. F. Female
837 preference for conspecific males based on olfactory cues in a Lake Malawi cichlid
838 fish. *Biol. Lett.* **1**, 411–414 (2005).
- 839 79. Milinski, M. Arms races, ornaments and fragrant genes: The dilemma of mate choice
840 in fishes. *Neurosci. Biobehav. Rev.* **46**, 567–572 (2014).
- 841 80. Wisenden, B. D. Olfactory assessment of predation risk in the aquatic environment.
842 *Philos. Trans. R. Soc. B Biol. Sci.* **355**, 1205–1208 (2000).
- 843 81. Johannesen, A., Dunn, A. M. & Morrell, L. J. Olfactory cue use by three-spined
844 sticklebacks foraging in turbid water: Prey detection or prey location? *Anim. Behav.*
845 **84**, 151–158 (2012).
- 846 82. Feuk, L., Carson, A. R. & Scherer, S. W. Structural variation in the human genome.
847 *Nat. Rev. Genet.* **7**, 85–97 (2006).
- 848 83. Inoue, N., Nishizumi, H., Naritsuka, H., Kiyonari, H. & Sakano, H. Sema7A/PlxnCl
849 signaling triggers activity-dependent olfactory synapse formation. *Nat. Commun.* **9**, 1–
850 11 (2018).
- 851 84. Milinski, M. *et al.* Mate choice decisions of stickleback females predictably modified by
852 MHC peptide ligands. (2005).
- 853 85. Andreou, D., Eizaguirre, C., Boehm, T. & Milinski, M. Mate choice in sticklebacks
854 reveals that immunogenes can drive ecological speciation. *Behav. Ecol.* **28**, 953–961
855 (2017).
- 856 86. Ohashi, J. *et al.* Extended linkage disequilibrium surrounding the hemoglobin E
857 variant due to malarial selection. *Am. J. Hum. Genet.* **74**, 1198–1208 (2004).
- 858 87. Lenz, T. L., Spirin, V., Jordan, D. M. & Sunyaev, S. R. Excess of Deleterious
859 Mutations around HLA Genes Reveals Evolutionary Cost of Balancing Selection. **33**,
860 2555–2564 (2016).
- 861 88. Alonso, S., López, S., Izagirre, N. & De La Rúa, C. Overdominance in the human
862 genome and olfactory receptor activity. *Mol. Biol. Evol.* **25**, 997–1001 (2008).
- 863 89. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: A flexible trimmer for Illumina
864 sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
- 865 90. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-
866 MEM. *ArXiv* 1–3 (2013).

- 867 91. McKenna, A. *et al.* The genome analysis toolkit: A MapReduce framework for
868 analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303
869 (2010).
- 870 92. Depristo, M. A. *et al.* A framework for variation discovery and genotyping using next-
871 generation DNA sequencing data. *Nat. Genet.* **43**, 491–501 (2011).
- 872 93. Delaneau, O., Howie, B., Cox, A. J., Zagury, J. F. & Marchini, J. Haplotype estimation
873 using sequencing reads. *Am. J. Hum. Genet.* **93**, 687–696 (2013).
- 874 94. King, T., Butcher, S. & Zalewski, L. *Apocrita - High Performance Computing Cluster*
875 *For Queen Mary University Of London.* (2017). doi:10.5281/ZENODO.438045
- 876 95. R Development Core Team. R: A language and environment for statistical computing.
877 R Foundation for Statistical Computing, Vienna, Austria. URL [http://www.R-](http://www.R-project.org/)
878 [project.org/](http://www.R-project.org/). (2019). doi:10.1007/978-3-540-74686-7
- 879 96. Pfeifer, B., Wittelsbrger, U., Ramos-Onsins, S. E. & Lercher, M. J. PopGenome: An
880 efficient swiss army knife for population genomic analyses in R. *Mol. Biol. Evol.* **31**,
881 1929–1936 (2014).
- 882 97. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 1–14
883 (2016).
- 884 98. Cabrera, C. P. *et al.* Uncovering networks from genome-wide association studies via
885 circular genomic permutation. *G3 Genes, Genomes, Genet.* **2**, 1067–1075 (2012).
- 886 99. J, A. A. and R. topGO: Enrichment analysis for Gene Ontology. *Bioconductor R*
887 package version 2.22.0 (2010).
888 doi:<http://www.bioconductor.org/packages/2.11/bioc/html/topGO.html>
- 889 100. Huber, W. *et al.* Orchestrating high-throughput genomic analysis with Bioconductor.
890 *Nat. Methods* **12**, 115–121 (2015).
- 891 101. Bates, D., Mchler, M., Bolker, B. M. & Walker, S. C. Fitting linear mixed-effects
892 models using lme4. *J. Stat. Softw.* **67**, 1–48 (2015).
- 893 102. Kuznetsova, A., Brockhoff, P. B. & Christensen, R. H. B. lmerTest Package: Tests in
894 Linear Mixed Effects Models. *J. Stat. Softw.* **82**, 1–26 (2017).
- 895 103. Chang, C. C. *et al.* Second-generation PLINK: Rising to the challenge of larger and
896 richer datasets. *Gigascience* **4**, 1–16 (2015).
- 897 104. Haller, B. C. & Messer, P. W. Evolutionary Modeling in SLiM 3 for Beginners. *Mol.*

898 *Biol. Evol.* **36**, 1101–1109 (2019).

899

Table 1. Distribution of candidate windows and genes across populations and population-pairs. Number of candidate windows are shown in bold. Populations and population-pairs are organised from most F_{ST} differentiated [left], to least F_{ST} differentiated [right].

Level of analysis	Population										Non-random Distribution		
	CA_L	CA_R	G1_L	G1_R	NO_L	NO_R	G2_L	G2_R	US_L	US_R	X^2	d.f.	P
Population	510	1623	250	1618	142	733	118	1211	80	127	5495.5	9	< 0.001
	145	341	86	241	43	255	47	163	32	52	745.6	9	< 0.001
Population-Pair	CA	G1	NO	G2	US						3033.1	4	< 0.001
	362	230	152	181	46						274.8	4	< 0.001

Table 2. Distribution of candidate windows and genes across the genome for both levels of analysis. Number of candidate windows are shown in bold.

Level of analysis	Chromosome																					Non-random distribution		
	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	XIII	XIV	XV	XVI	XVII	XVIII	XX	XXI	chrUn	X^2	d.f.	P
Population	194	126	95	1099	111	7	689	262	272	31	413	128	104	24	27	129	20	41	223	2062	355	15028	20	< 0.001
	66	30	32	212	26	5	187	89	92	5	92	45	54	15	6	47	11	15	95	189	92	1176.8	20	< 0.001
Population-Pair	195	82	26	690	63	8	472	183	103	18	331	181	69	13	21	83	9	23	276	2726	281	24767	20	< 0.001
	64	21	14	135	18	8	117	60	33	7	82	52	38	4	3	27	3	6	60	157	62	878.9	20	< 0.001

Figure Legends

Figure 1. Distribution of candidate windows and confounding factors across the three-spined stickleback genome for the Germany 1 (G1) sampling location. Tajima's D and θ_w were averaged across overlapping 7.5kb windows (light blue dots). Windows with Tajima's D above 2.0 and θ_w two standard deviations above the mean are highlighted with different colours. Blue circles represent candidate windows with no confounding factors, red circles may be confounded by associated overdominance, and green circles in the population-pair panel may have arisen from population structure. The grey central line represents the smoothed averages across windows. The sex chromosome (LG XIX) is not shown here. All 20 autosomes (LGI-LGXXI) and the unmapped chromosome (chrUn) are shown on the x-axis in ascending order. All other sampling locations are shown in supplementary fig. S1.

Figure 2. Enrichment of functional sites in candidate windows at the population level. P -values were $-\log$ transformed. The dashed grey line represents the significance threshold for the one-tailed P -values indicating significant (above the line, $P < 0.05$) enrichment of SNP categories in candidate windows in comparison to the genomic baseline. Populations are organised from most F_{ST} differentiated pair (CA) to least F_{ST} differentiated pair (US).

Figure 3. Distribution of candidate genes under balancing selection in *G. aculeatus*. The number of candidate genes identified in every possible combination of (A) population-pairs and (B) populations. Candidate genes private to a single population-pair (A) or population (B) are highlighted in different shades of orange. All combinations of populations that share a single candidate gene are not included in (B) due to space restrictions. Population-pairs and populations are organised from least F_{ST} divergent (US) to most F_{ST} divergent (CA) and are coloured by population or pair specific colours (A-B). (C) The distribution of private candidate genes along the genome identified in a single population-pair or population. Bars are coloured by the same shade of orange from A-B.

Figure 4. Tajima's D of candidate genes with exceptional signals of balancing selection. Each point represents a non-overlapping 2kb window, which are representative of the genomic regions surrounding: (A) *ENSGACG00000011738*, *XCR1A.1*, and *FYCO1A*; (B) *NXPE3*; (C) *Plxnc1*. The transcript and strand of each gene is represented by the arrows at the top of each plot. Populations are coloured by a population-pair specific colour, and dashed lines indicate river populations, and solid lines represent lake populations.

Figure 5. Characteristics of Candidate Windows. (A) ΔR^2 represents the average LD in candidate windows minus the genomic baseline for all populations. Dashed lines indicate river, and solid lines indicate lake populations. Populations are coloured by a population-pair specific colour. Grey box represents the two 10kb windows where LD was not significantly higher

around candidate windows. All data points are plotted to show the extent of variation around the mean. (B) Longer candidate windows (i.e., more contiguous windows) had more extended LD than smaller windows in all 10kb bins up to 90kb away. Each 10kb bin represents the pairwise distance between SNPs used in the R^2 calculations. Model significance is indicated below the pairwise bin name. Boxplots are ordered from smallest to largest physical candidate window length from left to right. Missing increments in the size of candidate windows indicates no candidate window of that length was identified. (C) In general, F_{ST} is elevated in candidate windows (grey boxes) when compared to the genomic baseline (white boxes) for both F_{ST} outlier windows (dashed line boxes) and non- F_{ST} outlier windows (solid line boxes). Population colours are retained from (A).

Figure 6. Estimation of population size history inferred using MSMC. For each population, 4 diploid genomes were used to infer effective population history. Populations are split by ecotype, with the [left] panel containing lake populations represented by solid lines, and the [right] panel containing river populations represented by dashed lines. Populations are coloured by a population-pair specific colour.

Figures

Figure 1

Distribution of candidate windows and confounding factors across the three-spined stickleback genome for the Germany 1 (G1) sampling location. Tajima's D and θ_w were averaged across overlapping 7.5kb windows (light blue dots). Windows with Tajima's D above 2.0 and θ_w two standard deviations above the mean are highlighted with different colours. Blue circles represent candidate windows with no confounding factors, red circles may be confounded by associated overdominance, and green circles in the population-pair panel may have arisen from population structure. The grey central line represents the smoothed averages across windows. The sex chromosome (LG XIX) is not shown here. All 20 autosomes (LGI-LGXXI) and the unmapped chromosome (chrUn) are shown on the x-axis in ascending order. All other sampling locations are shown in supplementary fig. S1.

Figure 2

Enrichment of functional sites in candidate windows at the population level. P-values were -log transformed. The dashed grey line represents the significance threshold for the one-tailed P-values indicating significant (above the line, $P < 0.05$) enrichment of SNP categories in candidate windows in comparison to the genomic baseline. Populations are organised from most FST differentiated pair (CA) to least FST differentiated pair (US).

Figure 3

Distribution of candidate genes under balancing selection in *G. aculeatus*. The number of candidate genes identified in every possible combination of (A) population-pairs and (B) populations. Candidate genes private to a single population-pair (A) or population (B) are highlighted in different shades of orange. All combinations of populations that share a single candidate gene are not included in (B) due to space restrictions. Population-pairs and populations are organised from least FST divergent (US) to most FST divergent (CA) and are coloured by population or pair specific colours (A-B). (C) The distribution of private candidate genes along the genome identified in a single population-pair or population. Bars are coloured by the same shade of orange from A-B.

Figure 4

Tajima's D of candidate genes with exceptional signals of balancing selection. Each point represents a non-overlapping 2kb window, which are representative of the genomic regions surrounding: (A) ENSGACG00000011738, XCR1A.1, and FYCO1A; (B) NXPE3; (C) Plxnc1. The transcript and strand of each gene is represented by the arrows at the top of each plot. Populations are coloured by a population-pair specific colour, and dashed lines indicate river populations, and solid lines represent lake populations.

Figure 5

Characteristics of Candidate Windows. (A) ΔR^2 represents the average LD in candidate windows minus the genomic baseline for all populations. Dashed lines indicate river, and solid lines indicate lake populations. Populations are coloured by a population-pair specific colour. Grey box represents the two 10kb windows where LD was not significantly higher around candidate windows. All data points are plotted to show the extent of variation around the mean. (B) Longer candidate windows (i.e., more contiguous windows) had more extended LD than smaller windows in all 10kb bins up to 90kb away. Each 10kb bin represents the pairwise distance between SNPs used in the R^2 calculations. Model significance is indicated below the pairwise bin name. Boxplots are ordered from smallest to largest physical candidate window length from left to right. Missing increments in the size of candidate windows indicates no candidate window of that length was identified. (C) In general, F_{ST} is elevated in candidate windows (grey boxes) when compared to the genomic baseline (white boxes) for both F_{ST} outlier windows (dashed line boxes) and non- F_{ST} outlier windows (solid line boxes). Population colours are retained from (A).

Figure 6

Estimation of population size history inferred using MSMC. For each population, 4 diploid genomes were used to infer effective population history. Populations are split by ecotype, with the [left] panel containing lake populations represented by solid lines, and the [right] panel containing river populations represented by dashed lines. Populations are coloured by a population-pair specific colour.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [CBThorburn2021SupplementaryTables.xlsx](#)
- [CBThorburn2021SupplementaryFigures.pptx](#)