PNAS
www.pnas.org

Supplementary Information for

Fitness maps to a large-effect locus in introduced stickleback populations

Dolph Schluter, Kerry B. Marchinko, Matt E. Arnegard, Haili Zhang, Shannon D. Brady, Felicity C. Jones, Michael A. Bell, and David M. Kingsley.

Co-corresponding authors: Dolph Schluter and David M. Kingsley
schluter@zoology.ubc.ca, kingsley@stanford.edu

This PDF file includes:

Methods

Quality filtering of markers

We filtered the markers as follows. Invariant markers and those having a missing genotype in the wild progenitors of the cross were deleted, as were those on the mitochondrion or Y-chromosome. We removed markers having either a median GenomeStudio genotype accuracy score less than 0.30, more than 40% of normalized *R*-values (GenomeStudio signal intensity) below 0.10, or a median GenTrain score (GenomeStudio measure of cluster quality) less than 0.30. Markers whose *Theta* values (GenomeStudio measurement of the fraction of non-reference alleles per individual) were not distinguishable between the two F0 individuals (difference < 0.20) were also filtered. We converted all genotype calls having a normalized *R*-value less than 0.10 to missing at remaining markers.

We used a custom script in R 3.2.3 (1) to standardize genotype calls between the two batches and obtain genotype probabilities for all F3 individuals. We applied adaptive kernel density estimation (2) to the frequency distribution of normalized *Theta* values for each marker using the *akj* function in the *quantreg* package (3). Bandwidth was chosen for each marker to yield a density with three genotype modes (two modes were fitted in the case of sex-linked markers). Markers unlinked to sex were dropped if three clusters were not obtained, as confirmed by visual inspection of scatter plots of GenomeStudio *R*-values against *Theta*. Normalized *Theta* values corresponding to fitted density minima were initially used to demarcate genotypes. We then fitted a Gaussian mixture model to the normalized *Theta* values to reassign individuals to genotypes based on their posterior probabilities of genotype cluster membership using the *me* function of the *mclust* package (4). Markers were dropped if genotype probabilities were less than 0.90 in more than 40% of genotype calls, and if posterior probability of assignment was less than 0.90 for all genotypes assigned to any given cluster. Four hundred fifty-eight markers remained for subsequent analyses after filtering.

Sex of individuals

We confirmed sex of putative F2 adult females and estimated sex of F3 juveniles using genotypes at 4 sex-linked markers: CH213-119K16:14070|gg1, CH213-119K16:207645|gg4, chrXIX:14650559|HFx055, and chrXIX:8190806|SNP1813 (5). Individuals homozygous at all genotyped markers were scored as female, whereas those heterozygous at all genotyped markers were called male. Individuals with a mixture of homozygous and heterozygous genotypes were scored as uncertain.

Parentage analysis

We carried out parentage assignment of F2 females to F1 parents, and of F3 offspring to F2 mothers, using the MasterBayes package in R (6) after removing the four sex-linked markers. We included only F2 individuals whose posterior probability of assignment to F1 parents was at least 0.99. Four F2 females having posterior probability of assignment to F1's less than 0.60 were removed from further analysis, leaving 224 F2 females for analysis. When assigning F3s to F2 parents we assumed that 10 potential F2

mothers and 200 potential F2 fathers were missing from the data set. The number of missing fathers of F3's is large because none were genotyped. The number of missing mothers of F3's is known to be small because of the high number of surviving F2's captured. Doubling the number of potential missing mothers did not change the results. Of 500 F3 individuals genotyped, 12 were assigned to missing mothers and were deleted from the data set. 451 were assigned to F2 female parents with probability 0.90 or higher, and 474 were assigned with probability greater than 0.60. We used the more lenient cutoff in subsequent analyses, but the higher cutoff gave the same results.

Linkage mapping and QTL analysis

In each F1 × F1 family we compared the frequency distribution of F2 genotypes at every marker to the random expectation (1:2:1 or 1:1 ratio, depending on bi-allelic marker state) using the Pearson χ^2 statistic. Observed χ^2 values and their degrees of freedom were summed across the 6 families and used to calculate a *P*-value for every marker. Markers departing from the random expectation with a Bonferroni corrected *P*-value < 0.01 were dropped, as were markers with unknown phase. We also removed sex-linked markers at this stage, since we used only F2 females, leaving 400 markers for mapping.

We used the cross pollinator option in JoinMap v.3.0 (39) to create a linkage map from the F2 cross. First, we analyzed each F1 × F1 family separately to obtain recombination frequencies (and associated LODs) between all pairs of markers. A single data file was then produced by concatenating the files of recombination frequencies obtained from the separate families. Finally, this concatenated file was used to estimate the joint linkage map in JoinMap using calculation options described in (7). Twenty-one linkage groups were identified at a LOD grouping threshold of 6.0, corresponding to the 21 chromosomes in the stickleback genome (41).

We used *R/qtl* to perform QTL mapping. F1 × F1 family identity was a covariate in all analyses (8). We conducted 10,000 permutations per trait to determine the LOD threshold (3.65) corresponding to a genome-wide significance level of $\alpha = 0.05$. For every QTL, we estimated the position of the peak LOD score in cM with a 1.5-LOD confidence interval to either side of the maximum (9). A linear model (*lm*) in R was used to estimate the percent of phenotypic variance explained by a QTL, fitting each trait to the genotype probabilities at the marker corresponding to the highest LOD score (hereafter, the peak marker) extracted using *pull.genoprob* in *R/qtl*. Linear models used sequential sums of squares and entered family and other covariates before genotype in model formulas. Model fits were visualized using conditional plots in *visreg* in R (10).

Lateral plate morph was mapped as a quantitative variable, with 0 corresponding to low-plated, 1 to high-plated, and 0.5 to partially plated. F2 female body size (standard length) and number of offspring were analyzed untransformed. Body size showed positive skew, but a log transformation did not change the results. A single body size outlier (an F2 female with standard length 2.55 cm) was removed from all analyses including this trait. Number of offspring was skewed and heteroscedastic, as expected for count data, but fitting reproductive success to genotype probabilities in R using *glm* with a quasipoisson error distribution and log link function did not alter the results. There was

a weak effect of F2 capture date on size measurements, but including it as a covariate did not change the results and we do not present them.

Eda genotyping of Loberg Lake fish

DNA from ethanol fixed fish was extracted from caudal fin clips as described (11). DNA from 10% buffered formalin fixed fish were prepared as follows: caudal fins were each cut using a new razor blade and the fin clip was washed in 900ul PBS (pH 7.4) (Gibco) on a nutating mixer at room temperature 3 times for 30min each. Fin clips were transferred using individual pipette tips to 250ul 10mM Tris-HCl (pH=8.5) (Qiagen) containing 0.5% Tween-20 (Sigma), sealed in microcentrifuge tubes with cap locker (E&K Scientific), and heated at 110°C for 20min. Tubes were then cooled to 55°C and incubated with 200 microgram per ml Proteinase K (Fermentas) overnight. After proteinase digestion, an equal volume (~250ul) of 10mM Tris-HCl (pH=8.5) containing 5% of Chelex-100 resin (Biorad) was added, and tubes were sealed with cap locker and heated at 110°C for another 20min. Samples were then centrifuged at 20,800g for 10min and supernatants were transferred to phase-lock gel (Fisher Scientific) loaded with 250ul chloroform (EMD) and mixed by inversion several times. Following centrifugation at 20800g for 10min at room temperature, 500 ul aqueous phase was transferred to new tubes containing 1000ul 100% ethanol and 50ul 3M buffered sodium acetate (Sigma), mixed by inversion, and incubated at -20°C overnight. DNA precipitates were collected by centrifugation, and pellets were washed with 70% ethanol, air dried for 10min, and resuspended in 30ul 10mM Tris-HCl (pH=8.5). All procedures for isolating DNA from formalin-fixed fish were performed in a sterile tissue culture hood with HEPA air filtration to minimize sample contamination.

DNA isolated from ethanol preserved fish was genotyped at the *Eda* locus by PCR amplification using primers GCCCTTCAATCCATCATCAG and TCCAATGATGTAAGAAGGCTCA, which produces a 668 bp product containing 3 SNPs that are characteristic for marine and freshwater populations as described (12). PCR reaction was carried out in 20ul using Phusion DNA polymerase (NEB) following manufacturer's instruction. Direct sequencing of PCR product was performed as described (12). DNA isolated from formalin preserved fish was genotyped at the *Eda* locus by PCR amplification using primers CCCAATTGTTCCAAAAATGAA and TAAAGAGCATTGGCCTCTGA, which generated a 108 bp product containing one SNP characteristic of marine or freshwater alleles. PCR was carried out using Restorase DNA polymerase (Sigma) following manufacturer's instructions. PCR products were cloned into Zero Blunt TOPO cloning kit (Invitrogen) and sequenced using M13F primer TGTAAAACGACGGCCAGT. *Eda* genotype calls were based on sequences from ≥5 colonies per PCR reaction. The number of individuals and genotypes from different sampling years are summarized in *SI Appendix*, Table S1.

Figures

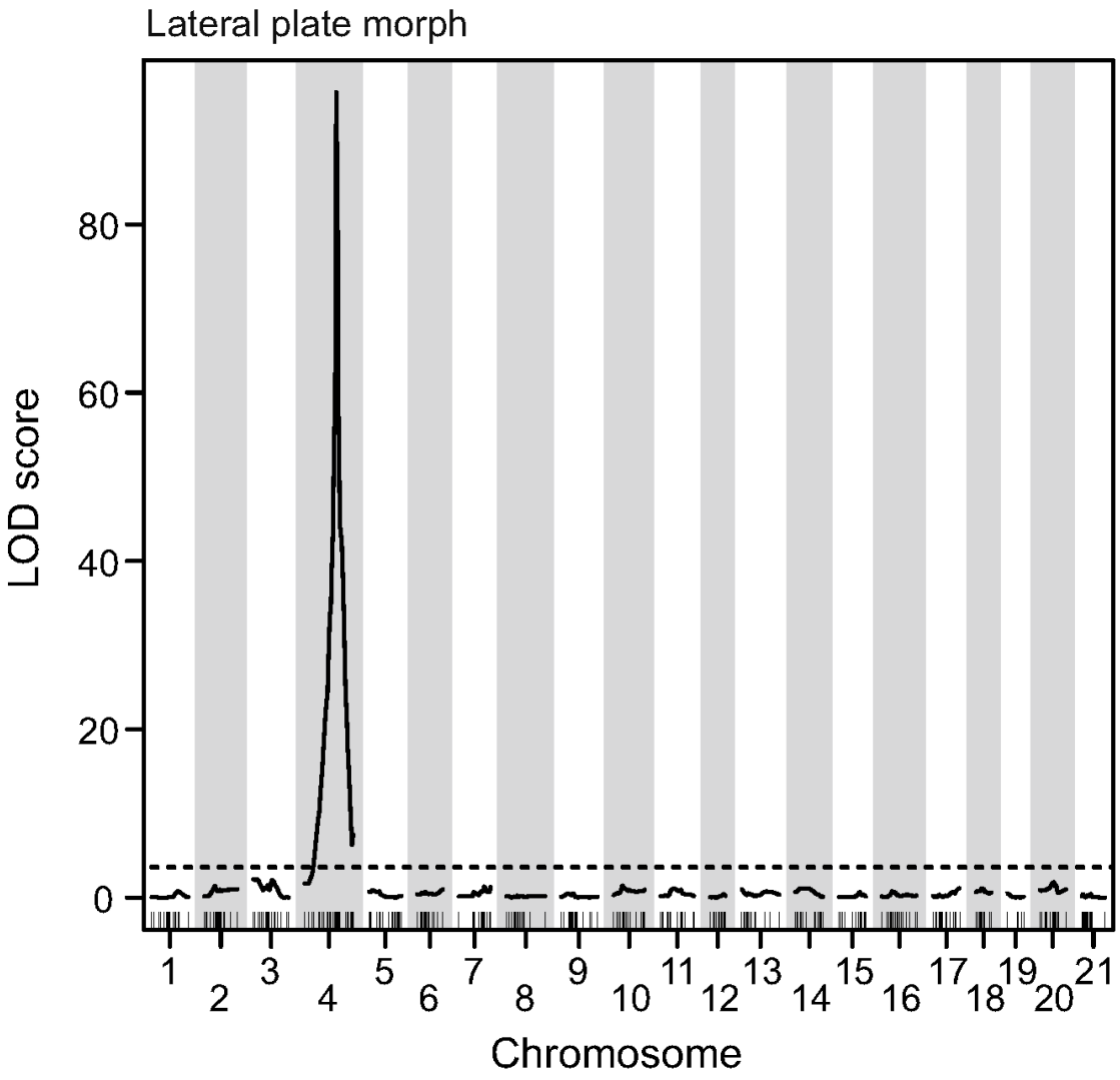

Figure S1. Quantitative trait locus (QTL) map for lateral plate morph of F2 females. The horizontal line indicates the LOD threshold of 3.65 corresponding to a genome-wide significance level of $\alpha = 0.05$.

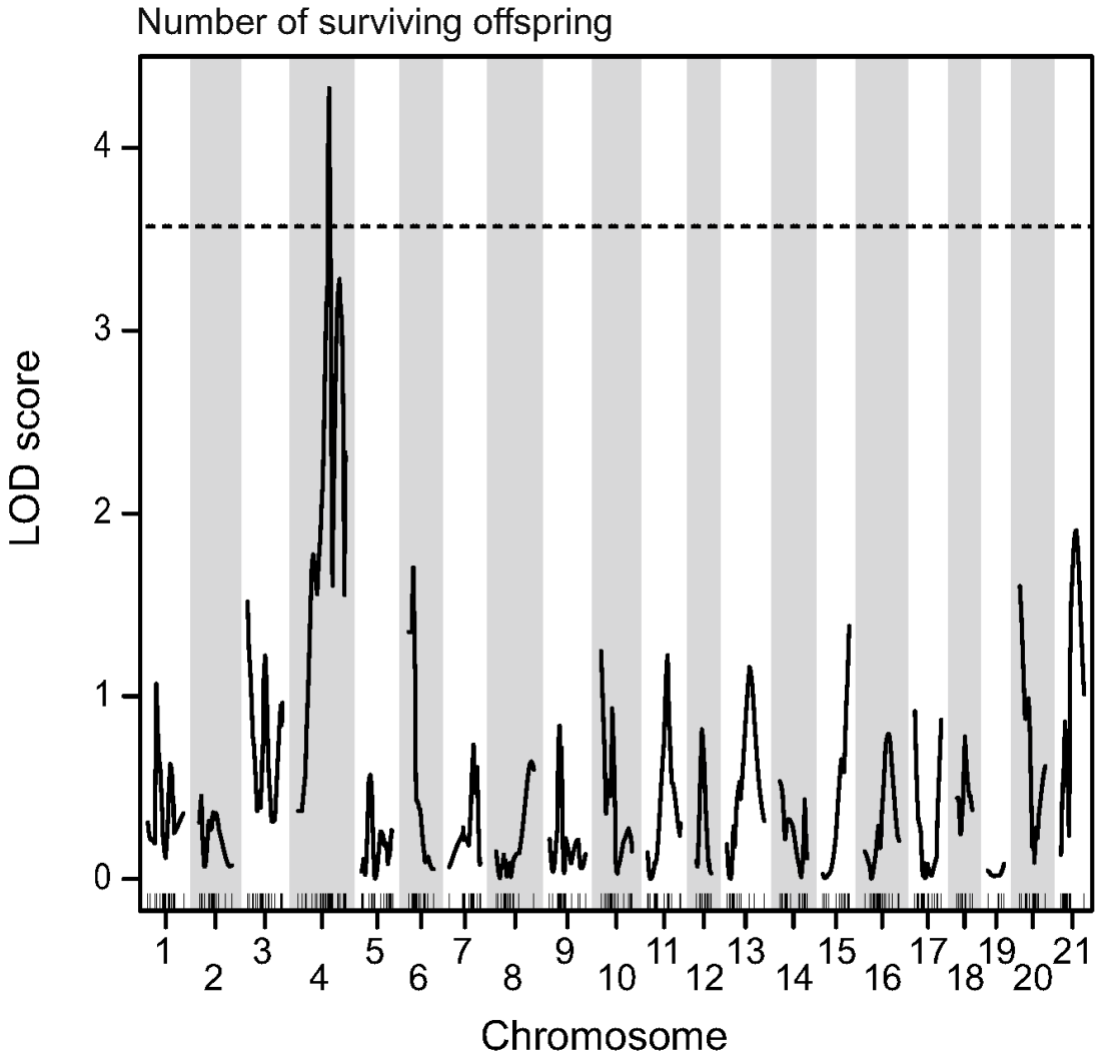

Figure S2. Quantitative trait locus (QTL) map for F2 female fitness measured as the number of surviving offspring. Family identity (unique combination of F1 parents) was included as a covariate. The horizontal line indicates the LOD threshold of 3.65, corresponding to a genome-wide significance level of $\alpha = 0.05$.

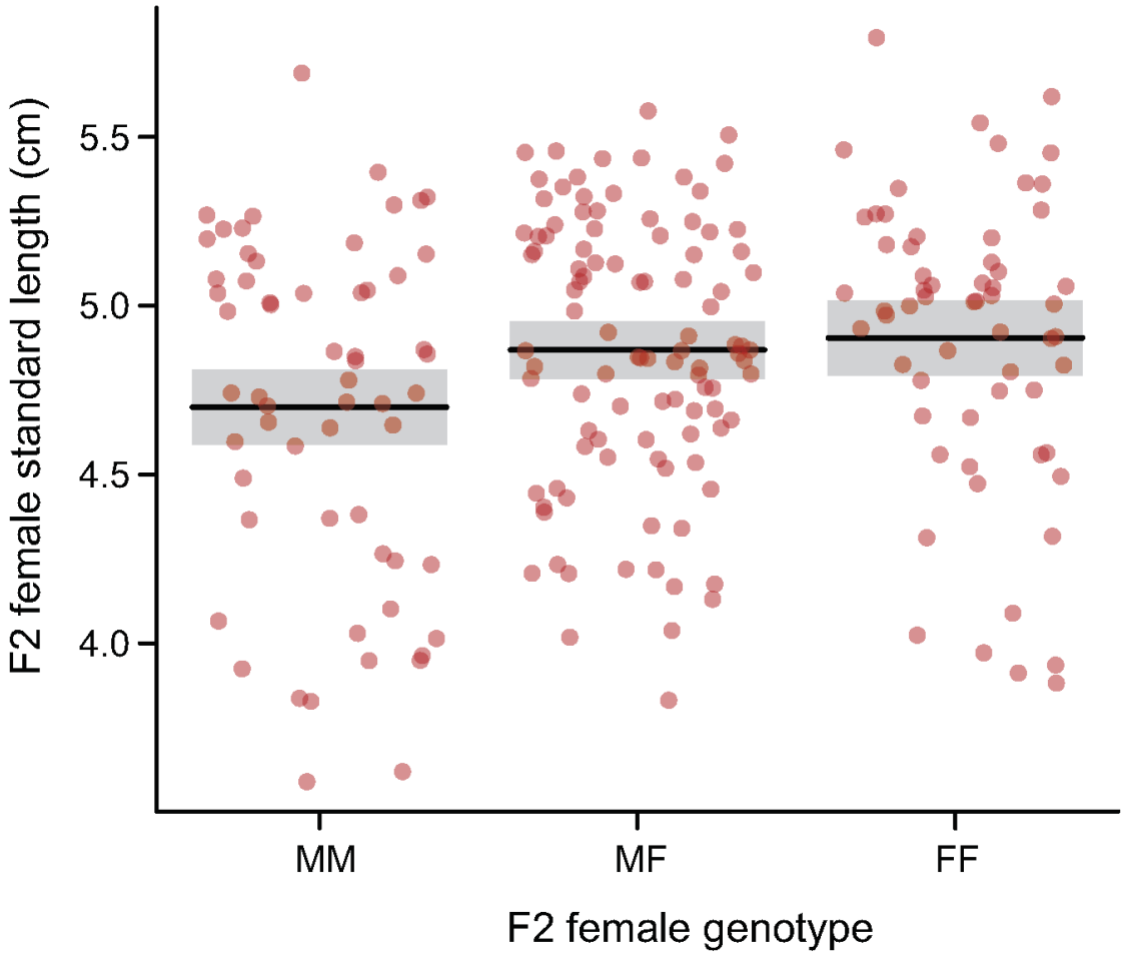

Figure S3. Body size (standard length) of F2 females differing in genotype at the peak marker for fitness (Figure 2). MM females are homozygous for the ancestral marine allele; FF females have two copies of the derived freshwater allele; and MF females are heterozygous. Horizontal line segments are means, and vertical span of shaded region is the 95% confidence interval for the mean, conditional on family identity (unique F1 × F1 parent combination). A single outlier of 2.55 cm standard length and 0 offspring was left out.

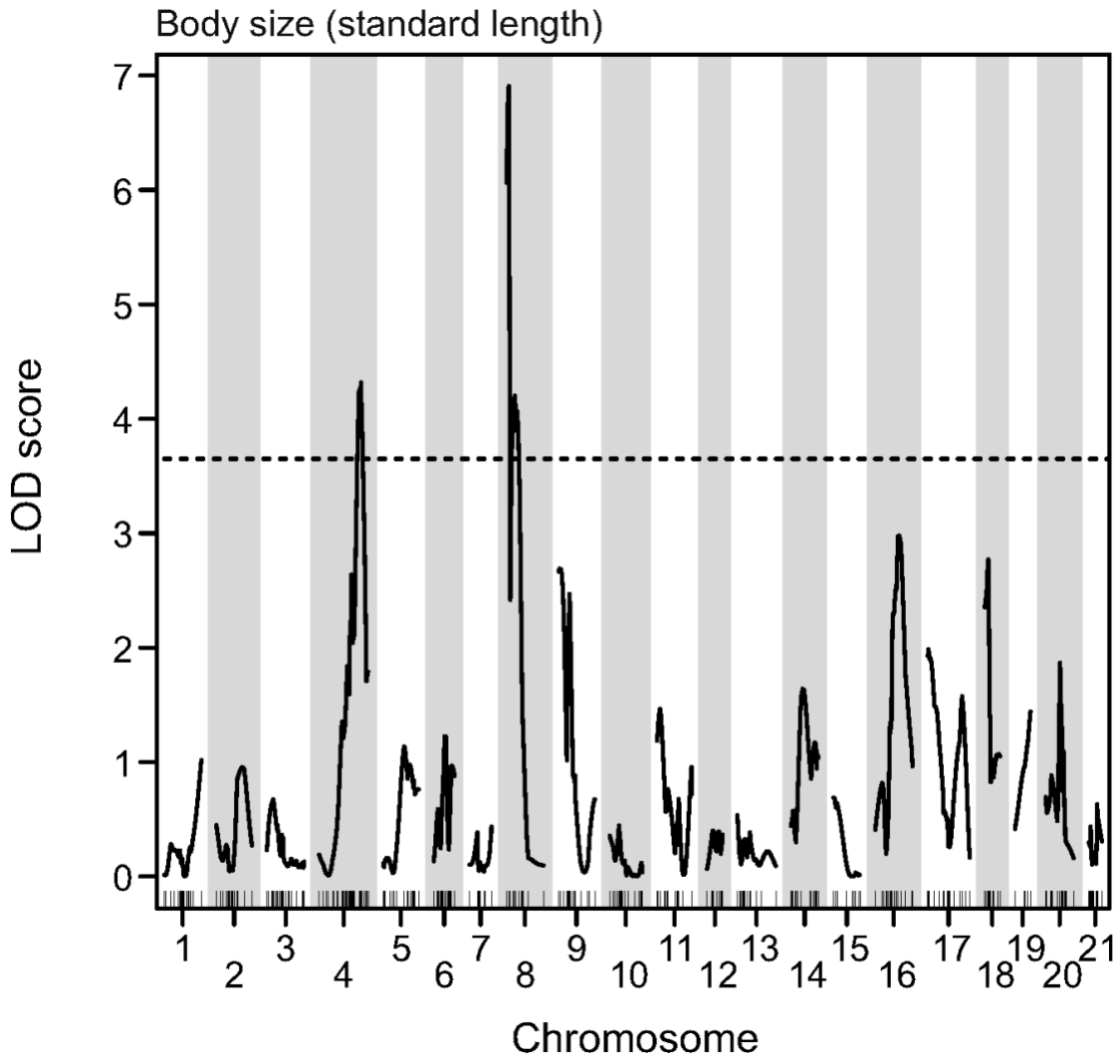

Figure S4. Quantitative trait locus (QTL) map of F2 female body size (standard length). Family identity was included as a covariate. A single outlier of 2.55 cm standard length and 0 offspring was left out. The horizontal dashed line indicates the threshold LOD score 3.65, corresponding to a genome-wise significance level of $\alpha = 0.05$.

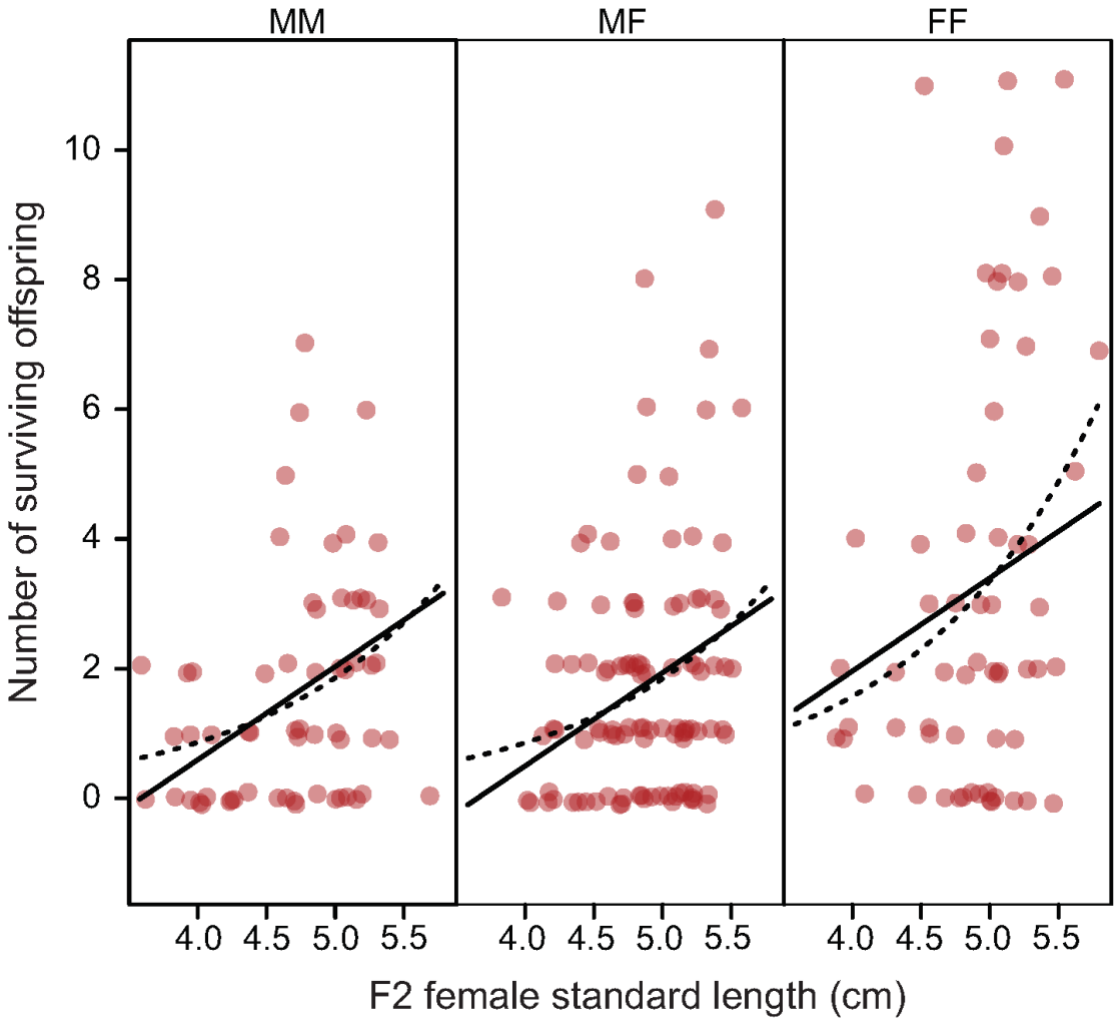

Figure S5. Numbers of surviving offspring of F2 mothers varying in body size (standard length) and genotype at the peak marker for fitness. Regression lines have the same slope but different intercepts. Points are displaced vertically by a small random amount to reduce overlap. Dashed line indicates the Poisson regression fit to the same data. A single outlier having a standard length of 2.55 cm and 0 offspring was left out of the analysis.

Tables

Table S1. The number of MM, MF, and FF genotypes at the *Eda* locus in Lake Loberg in each year, where M refers to the high-armor marine allele and F refers to the low-armor freshwater allele.

year	FF	MF	MM	No. F alleles	Total no. alleles
1992	8	29	10	45	94
1994	14	12	14	40	80
1996	18	19	4	55	82
1999	35	20	0	90	110
2001	70	12	1	152	166
2003	40	6	2	86	96
2005	77	11	0	165	176
2007	54	6	0	114	120
2008	68	11	0	147	158
2010	82	11	0	175	186

References

1. R Core Team (2017) R: A language and environment for statistical computing (R Foundation for Statistical Computing, Vienna, Austria).
2. Silverman BW (1986) *Density estimation for statistics and data analysis* (CRC press).
3. Koenker R (2015) quantreg: Quantile regression. R package version 5.19. http://CRAN.R-project.org/package=quantreg.
4. Fraley C, Raftery AE (2002) Model-based clustering, discriminant analysis, and density estimation. *J Am Stat Assoc* 97:611-631.
5. Arnegard ME, *et al.* (2014) Genetics of ecological divergence during speciation. *Nature* 511:307-311.
6. Hadfield JD, Richardson DS, Burke T (2006) Towards unbiased parentage assignment: combining genetic, behavioural and spatial data in a Bayesian framework. *Mol Ecol* 15:3715-3730.
7. Van Ooijen JW, Voorrips RE (2001) *JoinMap® 3.0, Software for the calculation of genetic linkage maps* (Plant Research International, Wageningen, The Netherlands).
8. Broman KW, Sen S (2009) *A guide to QTL mapping with R/qtl* (Springer).
9. Manichaikul A, Dupuis J, Sen Ś, Broman KW (2006) Poor performance of bootstrap confidence intervals for the location of a quantitative trait locus. *Genetics* 174:481-489.
10. Breheny P, Burchett W (2017) visreg: visualization of regression models. *R package version 24-1 https://CRANR-projectorg/package=visreg*.
11. Peichel CL, *et al.* (2001) The genetic architecture of divergence between threespine stickleback species. *Nature* 414:901-905.
12. Colosimo PF, *et al.* (2005) Widespread parallel evolution in sticklebacks by repeated fixation of *Ectodysplasin* alleles. *Science* 307:1928-1933.