DNA fragility in the parallel evolution of pelvic reduction in stickleback fish

Kathleen T. Xie[1,2,], Guliang Wang[3], Abbey C. Thompson[1], Julia I. Wucherpfennig[1], Thomas E. Reimchen[4], Andrew D. C. MacColl[5], Dolph Schluter[6], Michael A. Bell[7], Karen M. Vasquez[3], and David M. Kingsley[1,2*]

[1]Stanford University School of Medicine, Stanford, CA, USA

[2]Howard Hughes Medical Institute, Stanford, CA, USA

[3]University of Texas at Austin, Austin, TX USA

[4]University of Victoria, Victoria, BC, Canada

[5]University of Nottingham, Nottingham, UK

[6]University of British Columbia, Vancouver, BC, Canada

[7]University of California Museum of Paleontology, Berkeley, CA, USA

[*] Correspondence: kingsley@stanford.edu

ONE SENTENCE SUMMARY:

DNA fragility and high mutation rates influence the genomic pathways of adaptive evolution.

ABSTRACT

Evolution generates a remarkable breadth of living forms, but many traits evolve repeatedly, by still poorly understood mechanisms. A classic example of repeated evolution is the loss of pelvic hindfins in stickleback fish (*Gasterosteus aculeatus*). Repeated pelvic loss maps to recurrent deletions of a pelvic enhancer of the *Pitx1* gene. Here, we identify molecular features contributing to these recurrent deletions. *Pitx1* enhancer sequences form alternative DNA structures in vitro, and increase double-strand breaks and deletions in vivo. Enhancer mutability depends on DNA replication direction and is caused by (TG)-dinucleotide repeats. Modeling shows that elevated mutation rates can influence evolution under demographic conditions relevant for sticklebacks and humans. DNA fragility may thus help explain why the same loci are often used repeatedly during parallel adaptive evolution.

MAIN TEXT

Many phenotypic traits evolve repeatedly in organisms adapting to similar environments, and studying these cases can reveal ecological and genetic factors shaping parallel evolution (*1, 2*). For example, loss of pelvic appendages has evolved repeatedly in mammals, amphibians, reptiles, and fishes. Marine stickleback fish (*Gasterosteus aculeatus*) develop a robust pelvic apparatus, whereas many freshwater populations have lost pelvic structures (*3*). Pelvic reduction is associated with particular ecological conditions, is likely adaptive, and maps to recurrent and independent deletions of a pelvic enhancer (*Pel*) upstream of the homeodomain transcription factor gene (*Pitx1*) that also show repeatable molecular signatures of positive selection (*4-7*). This unusual spectrum of regulatory deletions contrasts with the accumulation of single nucleotide changes in other studies (*8, 6, 9*), hinting that special DNA features may shape adaptive variation at the *Pitx1* locus (*6*).

Pel enhancer sequences show high predicted helical twist flexibility (*6*), a DNA feature associated with delayed replication and fragile site instability (*10*). To examine whether *Pel* forms alternative DNA structures in vitro, we used 2-dimensional electrophoresis to analyze distributions of plasmid topoisomers (*11*) (Fig. 1A). A control stickleback genomic region showed smooth curves characteristic of B-DNA (Fig. 1B). In contrast, *Pel* sequences from marine populations showed mobility shifts characteristic of alternative DNA structure formation (Fig. 1B). Structural transitions started at a negative superhelical density of $-\sigma = 0.043$ and changed apparent linking numbers by 10-16 helical turns, similar to shifts produced by Z-DNA (left-handed DNA, starting $-\sigma = 0.046$) occupying ~105-170 bp (*12, 13*). *Pel* sequences from pelvic-reduced populations did not show unusual electrophoretic transitions (Fig. 1B), suggesting that natural mutations remove sequences forming alternative DNA structures.

To test the effect of *Pel* sequences on chromosome stability in vivo, we measured the rate of DNA double-strand breaks in yeast artificial chromosomes (Fig. 2A). Constructs without added test regions broke at background rates of 3.37 breaks per 10^6 divisions (Fig. 2B), consistent with previous reports (*14*). Chromosomes containing marine *Pel* broke ~25-50 times more frequently (Fig. 2B), a rate even higher than previously analyzed human fragile sites (*14*). *Pel* from freshwater pelvic-reduced populations (but not freshwater pelvic-complete populations

(fig. S1)) broke at rates similar to the control (Fig. 2B), suggesting that natural *Pel* mutations remove breakage-prone regions.

Reverse complements of marine *Pel* broke ~10-20 times less frequently than identical sequences in the forward orientation (Fig. 2B). RNA transcription can influence fragile site breakage (*15*), but reversing transcription orientation of the nearby *URA3* marker did not significantly affect *Pel* fragility (Fig. 2C). In contrast, adding a replication origin on the opposite side of *Pel* did switch fragility, making the forward sequence stable, and the reverse complement fragile (Fig. 2C). Thus, *Pel* fragility is markedly dependent on DNA replication direction.

Pel contains abundant runs of alternating pyrimidine-purine repeats (Fig. 3A, file S1), which can adopt alternative structures like Z-DNA, previously associated with deletions in bacteria, mice, and humans (*16, 17*). Three stretches of ~15, ~20, and ~50 (TG)-dinucleotide repeats in marine *Pel* total ~170 bp (consistent with linking number changes seen in topoisomer assays above). TG-repeats alone induced mobility shifts in topoisomer assays (Fig. 3B) (*18*) and elevated chromosome breakage in yeast, with longer repeats stimulating more breaks (Fig. 3C). In contrast, both long and short versions of the reverse complement sequence (CA-repeats) were stable (Fig. 3C), recapitulating the orientation dependence of *Pel* fragility.

We also tested the effect of TG- and CA-repeats in mammalian COS-7 cells (Fig. 3D) (*19*). Dinucleotide repeats elevated mutation frequencies, with TG-repeats being more mutagenic than CA-repeats of comparable lengths, and longer repeats being more mutagenic than shorter repeats (Fig. 3E), consistent with results from yeast assays. Mutations stimulated by the most mutagenic sequence, $(TG)_{41}$, were predominantly >100 bp deletions that removed part or all of the repeat and adjacent reporter gene (Fig. 3F, fig. S2A). Approximately 70% of deletion junctions contained microhomologies and insertions (Fig. 3F, fig. S2A-B), consistent with error-prone microhomology-mediated end-joining repair, and similar to junctions seen in stickleback pelvic-reduction alleles (*6*) (Fig. 3A). Ligation-mediated PCR suggested that breaks initiated near the dinucleotide repeats (fig. S2C). Taken together, our results indicate that TG-repeats form alternative DNA structures in vitro and can recapitulate the high mutation rates, orientation-dependence, and propensity to stimulate breaks and deletions of the full *Pel* region.

To determine the orientation of *Pel* sequences relative to DNA replication in sticklebacks (Fig. 4A, fig. S3), we sequenced S- and G-phase cells from developing embryos and calculated

S:G read-depth ratios to determine replication timing (*20*). *Pel* is located in a timing transition region (Fig. 4B, fig. S4), consistent with unidirectional replication. The replication direction through *Pel* matches the fragile orientation (Fig. 4C), suggesting that *Pel* would form a TG-repeat-associated fragile site in vivo. Experimental CRISPR targeting confirmed that initiation of breaks in *Pel* were sufficient to trigger local DNA deletions and macroscopic loss of pelvic structures in genetic crosses (fig. S5).

Could elevated mutation rates contribute to reuse of *Pel* deletions in parallel evolution? Population genetic modeling indicates that new mutations occurring at the low rates of typical single nucleotide changes ($\sim 10^{-9}$) would rarely arise at a particular locus in postglacial stickleback populations, while mutations occurring at elevated rates ($\sim 10^{-5}$ for fragile sites) would arise often. When new mutations do occur, their subsequent fate is controlled by drift and selection (*21*). Neutral or small-effect point mutations will usually be lost or rise to fixation slowly, while deletions may cause larger phenotypic effects and can sweep if environmental conditions favor pelvic reduction (Fig. 4D, fig. S6, fig. S7). The combined effects on both the "arrival of the fittest" and the "survival of the fittest" may explain why recurrent *Pel* deletions are the predominant mechanism for evolving stickleback pelvic reduction. For other traits, ancient standing variants provide an alternative way to overcome the demographic constraints of waiting for *de novo* mutations in small populations, and can also lead to reuse of similar alleles in different populations (*22, 23*).

The demographic parameters typical of sticklebacks apply to many vertebrates evolving with small population sizes or facing rapid environmental changes. For example, migration of modern humans out of Africa occurred with relatively small populations adapting to new environments in 3,000 generations or less (*24*). Interestingly, nearly half of currently known mutations underlying adaptive traits in modern humans also appear to be produced by mechanisms with elevated mutation rates (table S1).

High mutation rates have been described at contingency loci in bacteria and other systems (*25-30*). Our studies add an important new example of DNA fragility contributing to repeated morphological evolution in vertebrates. Our data also highlight several mechanisms that could alter local mutation rates, including expansion/contraction of TG-repeats, changes in sequence orientation, or changes in DNA replication. Natural variation in such parameters may

affect the evolvability of different loci and the particular genetic paths likely to be taken when ecological conditions favor a given phenotype. The sequence features associated with DNA fragility in the *Pel* region are also found in thousands of other positions in stickleback and human genomes (fig. S8). Notably, TG-repeats are enriched in other loci that have undergone recurrent ecotypic deletions during marine-freshwater stickleback evolution (*31*) (table S2, fig. S9), and near DNA breakage sites in humans (fig. S10). As causative changes are identified for more phenotypic traits, it will be interesting to see the extent to which DNA fragility has influenced the genes and mutations that underlie evolutionary change in nature.

ACKNOWLEDGEMENTS

We thank V. Tien, J. Le, M. Yau, M. Thakur, A. Muralidharan, M. Whitlock, B. Belotserkovskii, R. Driscoll, K. Cimprich, J. Wang, S. Quake, and A. Casper for experimental assistance or advice; R. Daugherty, J. Rollins, B. Lohman, R. Mollenhauer, M. Reyes, and F. von Hippel for help with fieldwork; C. Freudenreich for yeast strains; Z. Weng and B. Carter for help with high-throughput sequencing and cell sorting. **Funding:** NIH grants 5P50HG2568 (D.M.K.), CA093729 (K.M.V.), 2T32GM007790 (J.I.W); NSF grant DEB0919184 (M.A.B.); NSF and Stanford CEHG Graduate Fellowships (K.T.X.), NIH Predoctoral Fellowship (A.C.T.); HHMI investigator (D.M.K.). **Author contributions:** K.T.X and D.M.K designed the study. K.T.X., G.W., A.C.T., and J.I.W. performed experiments. K.T.X, G.W., A.C.T., D.S., K.M.V., and D.M.K. analyzed data. T.E.R., A.D.C.M, D.S., and M.A.B. provided key populations and comments. K.T.X and D.M.K wrote the paper with input from all authors**. Competing interests:** None. **Data and materials availability**: GEO accession GSE121537.

FIGURE LEGENDS

Figure 1. Marine but not freshwater *Pel* alleles form alternative structures in vitro. (A) Two-dimensional electrophoresis of circular DNA topoisomers. A distribution of plasmid topoisomers is separated on an agarose gel; each topological class forms one spot. Canonical B-DNA forms a smooth distribution. Alternative structures cause mobility shifts. Distribution shifts at the linking number inducing alternative structure. Dagger, mobility shift. **(B)** *Pel* from marine and freshwater pelvic-reduced populations. Control, *AtpA1*.

Figure 2. Marine but not freshwater *Pel* alleles break at high rates in yeast, in an orientation-dependent fashion. (A) Test DNA is inserted in a yeast artificial chromosome between two selectable markers (*LEU2* and *URA3*) and downstream of a telomere seed site. Breakage results in loss of *URA3*. **(B)** Box-and-whisker plot of *Pel* breakage rates. Whisker ends indicate maximum and minimum of 6 fluctuation assays (10 cultures each). RC, reverse complement. *$p < 0.01$ (table S5). Population names, table S6. **(C)** Reversing replication

direction through the test region, but not *URA3* transcription direction, reverses orientation of fragility. ori, DNA replication origin.

Figure 3. (TG)-dinucleotide repeats recapitulate structure formation, high breakage rate, orientation-dependence, and deletion spectrum. (A) To-scale maps of *Pel* in different freshwater pelvic-reduced populations (table S6). Green, *Pel* sequence driving pelvis expression (*6*). Light-brown, TG-repeats. White boxes, DNA deletions in indicated populations. Blue, DNA remaining. Letters, microhomologies at deletion junctions. **(B)** Two-dimensional gel for $(TG)_{30}$. Dagger, mobility shift. **(C)** Yeast artificial chromosome breakage rates for TG- or CA-repeats of varying lengths. *$p < 0.01$ (table S5). **(D)** Reporter shuttle plasmid schematic. **(E)** Mammalian mutation frequencies. Error bars indicate SEM of 4-5 independent experiments. *$p < 0.05$ (Student's t-test). Dagger, deletions dominate mutation spectrum (fig. S2A). **(F)** To-scale map of $(TG)_{41}$-induced deletions in mammalian cells.

Figure 4. *Pel* is located in the breakage-prone orientation in sticklebacks, generating a fragile site likely to contribute to parallel evolution in natural populations. (A) Workflow for profiling genome-wide replication timing. **(B)** Stickleback chromosome VII replication timing. Red line, *Pel* locus, which is subtelomeric. Hash marks, reference genome assembly gap. **(C)** Diagrams of stable and fragile replication orientations. ori, origin bubble. Purple, newly synthesized leading strand. Pink, newly synthesized lagging strand. **(D)** Probability of at least one de novo mutation arising at a particular locus in 10,000 generations and eventually fixing, as a function of typical stickleback population sizes (N) and mutation rates (μ, grey bars) for single nucleotides (SNPs), copy number variants (CNVs), and fragile sites. De novo point mutations are unlikely to occur and fix in small vertebrate populations, even when conferring a selective advantage (s=0.01, modeled here). In contrast, mutations occurring at fragile sites are likely to arise and contribute to repeated evolution when conferring a selective advantage. For additional parameters, including neutrality (s=0), see fig. S6 and fig. S7.

REFERENCES AND NOTES

1. D. Schluter, E. A. Clifford, M. Nemethy, J. S. McKinnon, Parallel evolution and inheritance of quantitative traits. *Am Nat* **163**, 809-822 (2004).

2. D. L. Stern, V. Orgogozo, Is genetic evolution predictable? *Science* **323**, 746-751 (2009).

3. M. A. Bell, Interacting evolutionary constraints in pelvic reduction of threespine sticklebacks, Gasterosteus aculeatus (Pisces, Gasterosteidae). *Biol J Linn Soc* **31**, 347-382 (1987).

4. T. E. Reimchen, Spine deficiency and polymorphism in a population of Gasterosteus aculeatus: an adaptation to predators? *Can J Zool* **58**, 1232-1244 (1980).

5. M. A. Bell, G. Orti, J. A. Walker, J. P. Koenings, Evolution of pelvic reduction in threespine stickleback fish: a test of competing hypotheses. *Evolution* **47**, 906-914 (1993).

6. Y. F. Chan *et al.*, Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a Pitx1 enhancer. *Science* **327**, 302-305 (2010).

7. M. Karhunen, J. Merila, T. Leinonen, J. M. Cano, O. Ovaskainen, DRIFTSEL: an R package for detecting signals of natural selection in quantitative traits. *Mol Ecol Resour* **13**, 746-754 (2013).

8. B. Prud'homme *et al.*, Repeated morphological evolution through cis-regulatory changes in a pleiotropic gene. *Nature* **440**, 1050-1053 (2006).

9. D. L. Stern, N. Frankel, The structure and evolution of cis-regulatory regions: the shavenbaby story. *Philos T R Soc B* **368**, (2013).

10. R. G. Thys, C. E. Lehman, L. C. Pierce, Y. H. Wang, DNA secondary structure at chromosomal fragile sites in human disease. *Curr Genomics* **16**, 60-70 (2015).

11. R. Bowater, F. Aboul-Ela, D. M. Lilley, Two-dimensional gel electrophoresis of circular DNA topoisomers. *Methods Enzymol* **212**, 105-120 (1992).

12. A. Nordheim, A. Rich, The sequence (dC-dA)n X (dG-dT)n forms left-handed Z-DNA in negatively supercoiled plasmids. *Proc Natl Acad Sci USA* **80**, 1821-1825 (1983).

13. A. Rich, A. Nordheim, A. H. Wang, The chemistry and biology of left-handed Z-DNA. *Annu Rev Biochem* **53**, 791-846 (1984).

14. H. Zhang, C. H. Freudenreich, An AT-rich sequence in human common fragile site FRA16D causes fork stalling and chromosome breakage in S. cerevisiae. *Mol Cell* **27**, 367-379 (2007).

15. A. Helmrich, M. Ballarino, L. Tora, Collisions between replication and transcription complexes cause common fragile site instability at the longest human genes. *Mol Cell* **44**, 966-977 (2011).

16. G. Wang, L. A. Christensen, K. M. Vasquez, Z-DNA-forming sequences generate large-scale deletions in mammalian cells. *Proc Natl Acad Sci USA* **103**, 2677-2682 (2006).

17. G. Wang, S. Carbajal, J. Vijg, J. DiGiovanni, K. M. Vasquez, DNA structure-induced genomic instability in vivo. *J Natl Cancer Inst* **100**, 1815-1817 (2008).

18. H. Hamada, M. G. Petrino, T. Kakunaga, M. Seidman, B. D. Stollar, Characterization of genomic poly(dT-dG).poly(dC-dA) sequences: structure, organization, and conformation. *Mol Cell Biol* **4**, 2610-2621 (1984).

19. G. Wang, K. M. Vasquez, Naturally occurring H-DNA-forming sequences are mutagenic in mammalian cells. *Proc Natl Acad Sci USA* **101**, 13448-13453 (2004).

20. N. Rhind, D. M. Gilbert, DNA replication timing. *Cold Spring Harb Perspect Biol* **5**, a010132 (2013).

21. M. Kimura, On the probability of fixation of mutant genes in a population. *Genetics* **47**, 713-719 (1962).

22. P. F. Colosimo *et al.*, Widespread parallel evolution in sticklebacks by repeated fixation of Ectodysplasin alleles. *Science* **307**, 1928-1933 (2005).

23. R. D. Barrett, D. Schluter, Adaptation from standing genetic variation. *Trends Ecol Evol* **23**, 38-44 (2008).

24. 1000_Genomes_Project_Consortium, A global reference for human genetic variation. *Nature* **526**, 68-74 (2015).

25. R. Moxon, C. Bayliss, D. Hood, Bacterial contingency loci: the role of simple sequence DNA repeats in bacterial adaptation. *Annu Rev Genet* **40**, 307-333 (2006).

26. A. Stoltzfus, L. Y. Yampolsky, Climbing mount probable: mutation as a cause of nonrandomness in evolution. *J Hered* **100**, 637-647 (2009).

27. X. Du *et al.*, Potential non-B DNA regions in the human genome are associated with higher rates of nucleotide mutation and expression variation. *Nucleic Acids Res* **42**, 12367-12379 (2014).

28. S. C. Galen *et al.*, Contribution of a mutational hot spot to hemoglobin adaptation in high-altitude Andean house wrens. *Proc Natl Acad Sci USA* **112**, 13958-13963 (2015).

29. A. Bacolla, J. A. Tainer, K. M. Vasquez, D. N. Cooper, Translocation and deletion breakpoints in cancer genomes are associated with potential non-B DNA-forming sequences. *Nucleic Acids Res* **44**, 5673-5688 (2016).

30. A. D. Hargreaves *et al.*, Genome sequence of a diabetes-prone rodent reveals a mutation hotspot around the ParaHox gene cluster. *Proc Natl Acad Sci USA* **114**, 7677-7682 (2017).

31. C. B. Lowe *et al.*, Detecting copy number variation between groups of samples. *Genome Res* **28**, 256-265 (2018).

32. V. P. Schulz, V. A. Zakian, The Saccharomyces PIF1 DNA helicase inhibits telomere elongation and de novo telomere formation. *Cell* **76**, 145-155 (1994).

33. G. Wang, S. Gaddis, K. M. Vasquez, Methods to detect replication-dependent and replication-independent DNA structure-induced genetic instability. *Methods* **64**, 67-72 (2013).

34. K. H. Schmidt, V. Pennaneach, C. D. Putnam, R. D. Kolodner, Analysis of gross-chromosomal rearrangements in Saccharomyces cerevisiae. *Methods Enzymol* **409**, 462-476 (2006).

35. T. G. Montague, J. M. Cruz, J. A. Gagnon, G. M. Church, E. Valen, CHOPCHOP: a CRISPR/Cas9 and TALEN web tool for genome editing. *Nucleic Acids Res* **42**, W401-407 (2014).

36. A. N. Shah, C. F. Davey, A. C. Whitebirch, A. C. Miller, C. B. Moens, Rapid reverse genetic screening using CRISPR in zebrafish. *Nat Methods* **12**, 535-540 (2015).

37. H. Swarup, Stages in the development of the stickleback Gasterosteus aculeatus. *J Embryol Exp Morphol* **6**, 373-383 (1958).

38. T. Ryba *et al.*, Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types. *Genome Res* **20**, 761-770 (2010).

39. H. Li, R. Durbin, Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589-595 (2010).

40. N. Crosetto *et al.*, Nucleotide-resolution DNA double-strand break mapping by next-generation sequencing. *Nat Methods* **10**, 361-365 (2013).

41. A. Koren *et al.*, Differential relationship of DNA replication timing to different forms of human mutation and variation. *Am J Hum Genet* **91**, 1033-1040 (2012).

42. C. B. Lowe *et al.*, Three periods of regulatory innovation during vertebrate evolution. *Science* **333**, 1019-1024 (2011).

43. M. Lynch, Rate, molecular spectrum, and consequences of human mutation. *Proc Natl Acad Sci USA* **107**, 961-968 (2010).

44. C. D. Campbell, E. E. Eichler, Properties and rates of germline mutations in humans. *Trends Genet* **29**, 575-584 (2013).

45. M. A. Bell, S. A. Foster, *The evolutionary biology of the threespine stickleback*. (Oxford University Press, New York, 1994), pp. 571.

46. J. L. Rollins, Body size and growth rate divergence among populations of threespine stickleback (Gasterosteus aculeatus) in Cook Inlet, Alaska, USA. *Can J Zool* **95**, 877-884 (2017).

47. M. D. Shapiro *et al.*, Genetic and developmental basis of evolutionary pelvic reduction in threespine sticklebacks. *Nature* **428**, 717-723 (2004).

48. R. Frankham, Effective population size/adult population size ratios in wildlife: a review. *Genet Res* **66**, 96-107 (1995).

49. J. DeFaveri, J. Merila, Temporal stability of genetic variability and differentiation in the three-spined stickleback (Gasterosteus aculeatus). *PLoS One* **10**, e0123891 (2015).

50. A. Perez-Figueroa, C. Fernandez, R. Amaro, M. Hermida, E. San Miguel, Population structure and effective/census population size ratio in threatened three-spined stickleback populations from an isolated river basin in northwest Spain. *Genetica* **143**, 403-411 (2015).

51. R. D. Barrett, S. M. Rogers, D. Schluter, Natural selection on a major armor gene in threespine stickleback. *Science* **322**, 255-257 (2008).

52. G. Hunt, M. A. Bell, M. P. Travis, Evolution toward a new adaptive optimum: phenotypic evolution in a fossil stickleback lineage. *Evolution* **62**, 700-710 (2008).

53. M. Kimura, T. Ohta, The average number of generations until fixation of a mutant gene in a finite population. *Genetics* **61**, 763-771 (1969).

54. M. Kimura, The length of time required for a selectively neutral mutant to reach fixation through random frequency drift in a finite population. *Genet Res* **15**, 131-133 (1970).

55. H. Reyes-Centeno, Out of Africa and into Asia: fossil and genetic evidence on modern human origins and dispersals. *Quatern Int* **416**, 249-262 (2016).

56. A. Martin, V. Orgogozo, The loci of repeated evolution: a catalog of genetic hotspots of phenotypic variation. *Evolution* **67**, 1235-1250 (2013).

57. M. A. Bell, G. Orti, Pelvic reduction in threespine stickleback from Cook Inlet lakes: geographical distribution and intrapopulation variation. *Copeia*, 314-325 (1994).

58. I. Hiratani *et al.*, Global reorganization of replication domains during embryonic stem cell differentiation. *Plos Biol* **6**, e245 (2008).

59. J. D. McPhail, Ecology and evolution of sympatric sticklebacks (Gasterosteus): evidence for a species-pair in Paxton Lake, Texada Island, British Columbia. *Can J Zool* **70**, 361-369 (1992).

60. S. Saxonov, P. Berg, D. L. Brutlag, A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proc Natl Acad Sci USA* **103**, 1412-1417 (2006).

61. P. H. Sudmant *et al.*, An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75-81 (2015).

Figure 1

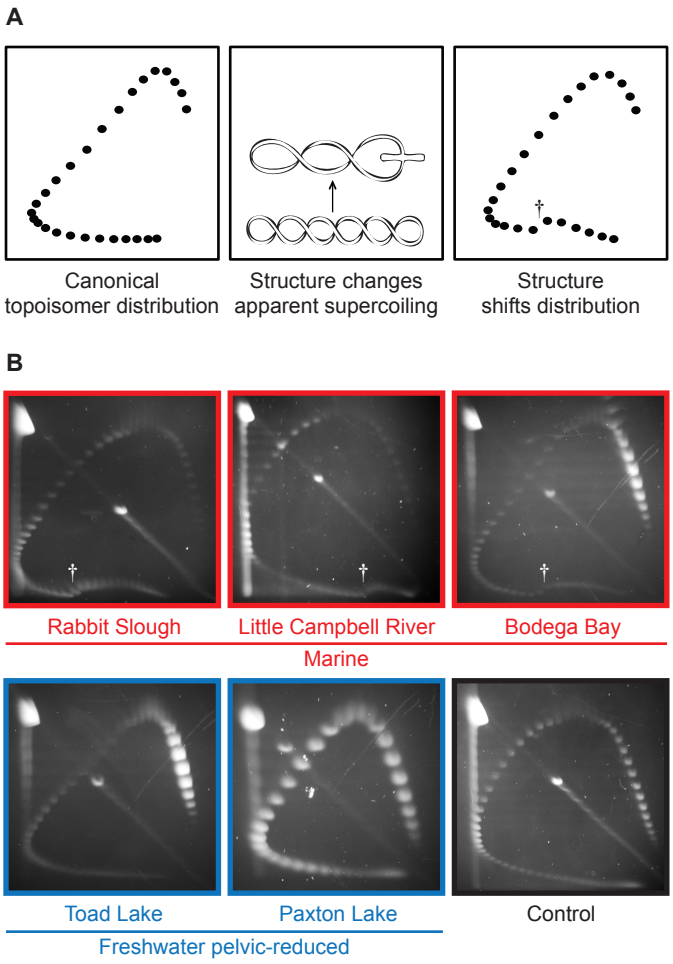

A
Canonical topoisomer distribution
Structure changes apparent supercoiling
Structure shifts distribution
B
Rabbit Slough
Little Campbell River
Bodega Bay
Marine
Toad Lake
Paxton Lake
Control
Freshwater pelvic-reduced

Figure 2

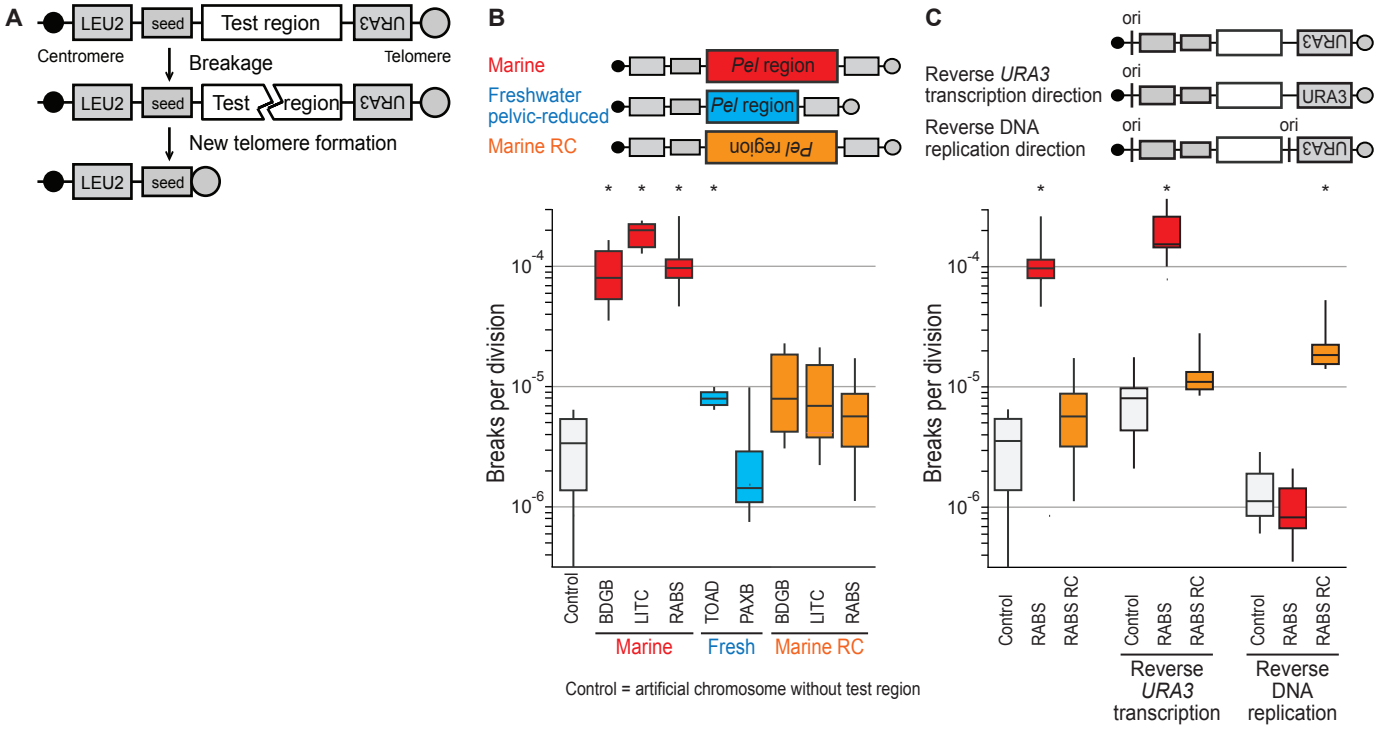

A
LEU2
seed
Test region
URA3
Centromere
Telomere
Breakage
Test
region
New telomere formation
B
Marine
Freshwater pelvic-reduced
Marine RC
Pel region
Breaks per division
Control
BDGB
LITC
RABS
TOAD
PAXB
Marine
Fresh
Marine RC
Control = artificial chromosome without test region
C
ori
Reverse URA3 transcription direction
Reverse DNA replication direction
URA3
RABS RC
Reverse URA3 transcription
Reverse DNA replication

Figure 3

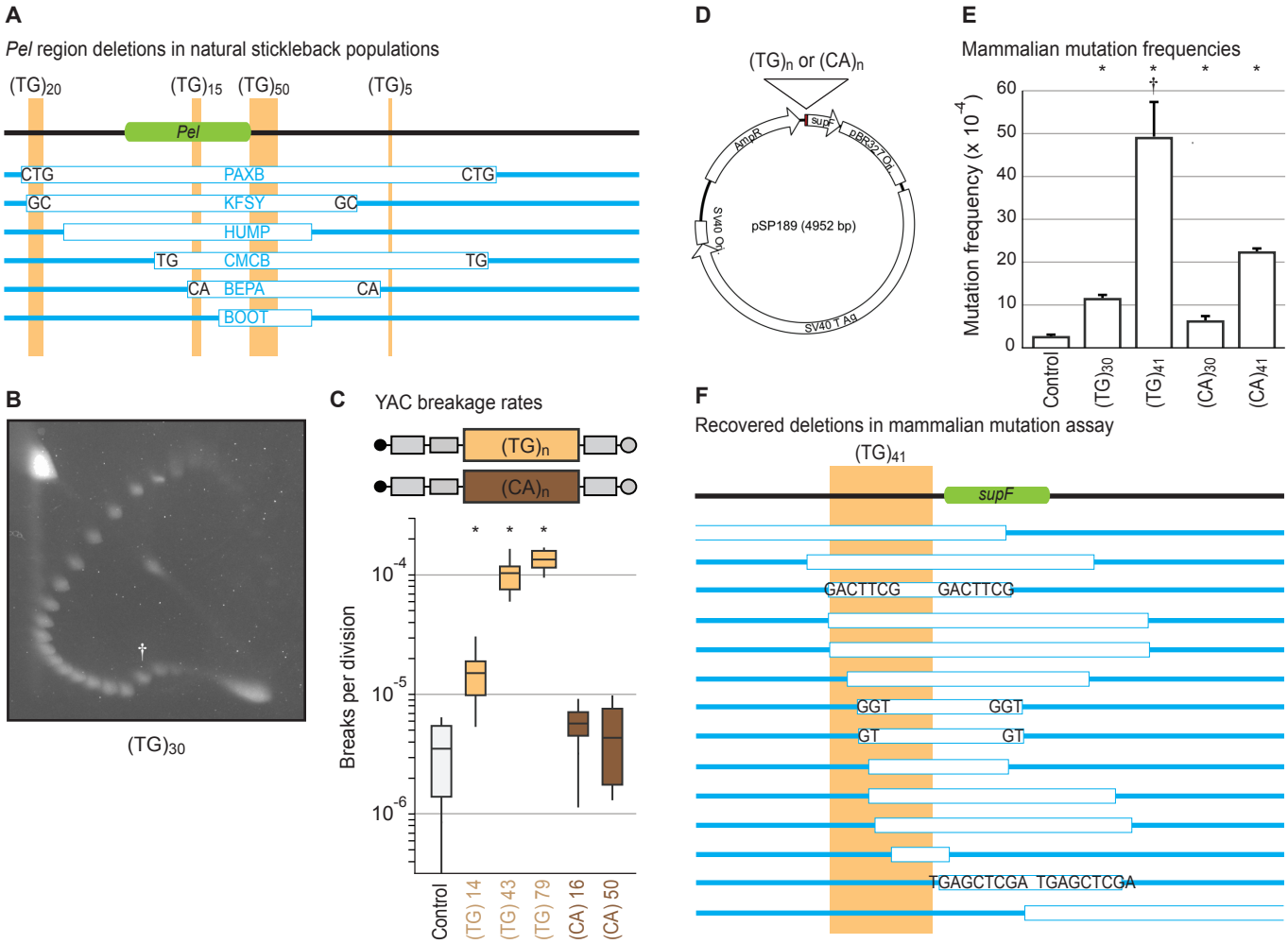
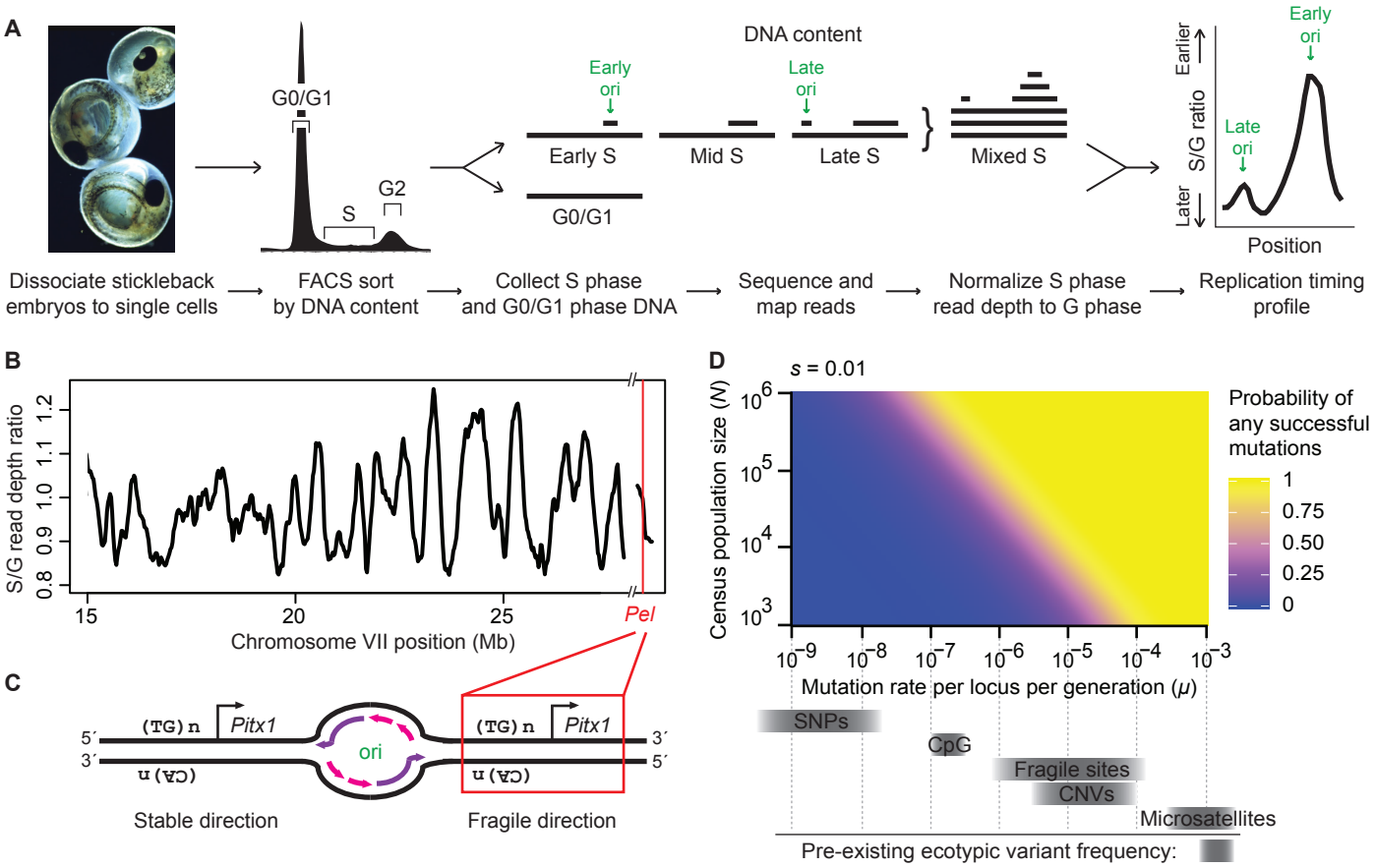

A
Pel region deletions in natural stickleback populations
(TG)20
(TG)15
(TG)50
(TG)5
Pel
CTG
PAXB
CTG
GC
KFSY
GC
HUMP
TG
CMCB
TG
CA
BEPA
CA
BOOT
B
(TG)30
C
YAC breakage rates
(TG)n
(CA)n
Breaks per division
10-4
10-5
10-6
Control
(TG) 14
(TG) 43
(TG) 79
(CA) 16
(CA) 50
D
(TG)n or (CA)n
AmpR
supF
pBR322 Ori
pSP189 (4952 bp)
SV40 Ori
SV40 T Ag
E
Mammalian mutation frequencies
Mutation frequency (x 10-4)
60
50
40
30
20
10
0
Control
(TG)30
(TG)41
(CA)30
(CA)41
F
Recovered deletions in mammalian mutation assay
(TG)41
supF
GACTTCG
GACTTCG
GGT
GGT
GT
GT
TGAGCTCGA
TGAGCTCGA

Figure 4


A
G0/G1
S
G2
DNA content
Early ori
Late ori
Early S
Mid S
Late S
Mixed S
G0/G1
Earlier
S/G ratio
Later
Position
Dissociate stickleback embryos to single cells
FACS sort by DNA content
Collect S phase and G0/G1 phase DNA
Sequence and map reads
Normalize S phase read depth to G phase
Replication timing profile
B
S/G read depth ratio
Chromosome VII position (Mb)
Pel
C
(TG)n
Pitx1
ori
Stable direction
Fragile direction
D
s = 0.01
Census population size (N)
Probability of any successful mutations
Mutation rate per locus per generation (μ)
SNPs
CpG
Fragile sites
CNVs
Microsatellites
Pre-existing ecotypic variant frequency:

Materials and Methods

Stickleback collection and care

A list of stickleback populations used and collected in this study is provided (table S6). Sticklebacks were captured using minnow traps or small minnow seines in less than 1 m deep water within 5 m of lakeshores or small coastal streams. Most marine sticklebacks were reproductive adults that were captured in freshwater, into which they ran to breed. All animal studies were performed in accordance with the recommendations in the Guide for the Care and Use of Laboratory Animals of the National Institutes of Health, using protocols approved by the Institutional Animal Care and Use Committee of Stanford University (IACUC protocol #13834), in animal facilities accredited by the Association for Assessment and Accreditation of Laboratory Animal Care International (AAALAC).

Primers

Primer sequences are as follows:

Pitx1-128,592.F 5′- CGCCGCTGCGTGAGAATATG
Pitx1-129,065.F 5′- TGACGCGGCGCTCCATCACCGAGCC
Pitx1-129,067.F 5′- ACGCGGCGCTCCATCACC
Pitx1-129,112.F 5′- GCTTGTAAAGAAGGGGAACCC
Pitx1-130,662.R 5′- CCTCAGATCTATCGCAGTAC
Pitx1-131,376.R 5′- CACAGCGAGCTGCTTTACGG
Pitx1-131,758.R 5′- TCTCTCAGCGGAGAAATCCG
Pitx1-132,268.R 5′- AGCTTCGTCACGCCACCTG
Atp1a1.F 5′- TGGAACGCTCTGGCCCGAAT
Atp1a1.R 5′- TGACGAAGAAAGCTGTGTGGCAT
T6-NsiI-URA3.F 5′- TTTTTTATGCATCGCGAGGCTGGATGGCCTTC
T6-NruI-URA3.R 5′- TTTTTTTCGCGATTACTTATAATACAGTTTTTTAG
ARSH4.F 5′- GATCGCCAACAAATACTACCT
ARSH4.R 5′- GGATCGCTTGCCTGTAACTT

Plasmid construction

Plasmids used for two-dimensional gel electrophoresis were constructed by TOPO cloning PCR products into pCR2.1-TOPO (Invitrogen). *Pitx1* enhancer region PCR primers used for RABS, LITC, BDGB, TOAD, and LSHP were Pitx1-129,065.F and Pitx1-131,758.R. Primers used for PAXB, ORPH, and BOUL were Pitx1-129,067.F and Pitx1-131,376.R. The control region used was a section of the *Atp1a1* gene, using primers Atp1a1.F and Atp1a1.R. PCR reactions were 200 uM dNTPs, 250 nM forward primer, 250 nM reverse primer, 50 mM Tris-HCl, 22 mM $(NH_4)_2SO_4$, 1.5 mM $MgCl_2$, and 0.1 u/uL Taq polymerase (New England Biolabs) or Expand High Fidelity Polymerase (Roche). High quality fresh genomic DNA preps were essential for successful PCR amplification. Plasmids used to construct yeast artificial chromosomes were derived from pVS20 (*32*), generously provided by Catherine Freudenreich. Because the stickleback *Pitx1* enhancer sequence contains many AatII recognition sites, AatII was not suitable for linearizing the plasmid. Thus, a NotI site was inserted in the AatII site of pVS20 to create pVS20N (such that the original GACGTC sequence was replaced by GACGTCAGCGGCCGCTGACGTC). To reverse the transcription direction of the *URA3* marker gene, the *URA3* insert was amplified by PCR

from pVS20N using the T6-NsiI-URA3.F and T6-NruI-URA3.R primers and digested with NsiI and NruI. This insert was then ligated into a vector prepared by digesting pVS20N with NsiI and NruI, to create the pVS20NR plasmid (reverse transcription empty vector). To add the replication origin, the ARSH4 origin was amplified by PCR from pRS316 using the ARSH4.F and ARSH4.R primers and inserted into pVS20N at the NruI site to create the pVS20N+ori plasmid (reverse replication empty vector). Stickleback *Pitx1* enhancer sequence was excised using EcoRI from the pCR2.1 constructs described above, and inserted into pVS20N, pVS20NR (reverse transcription), or pVS20N+ori (reverse replication), at the NsiI site, by blunt end cloning with Klenow, and transforming into SURE2 cells (Agilent). We call the "forward" orientation where the TG strand of the (TG)/(CA)-dinucleotide repeat stretches in the *Pitx1* enhancer region are on the G_4T_4 strand of the telomere seed sequence and the noncoding strand of the URA3 gene in pVS20N. $(TG)_{50}$ oligos were annealed to $(CA)_{50}$ oligos, and then cloned into the pVS20N NsiI site (Klenow blunted) to produce pVS20N $(TG)_n$ and pVS20N $(CA)_n$ plasmids, where n was of varying lengths determined by Sanger sequencing (Sequetech BDX chemistry). The lengths of the repeats stayed stable in SURE2 cells over many generations. The plasmids for mammalian mutation assays were constructed by cloning TG and CA-repeats into the *supF* mutation-reporter shuttle vector pSP189 as previously described (*19, 33*) at the EcoRI-XhoI restriction sites to create pSP189-$(TG)_{41}$, pSP189-$(CA)_{41}$, pSP189-$(TG)_{30}$, and pSP189-$(CA)_{30}$.

Two-dimensional gel electrophoresis of topoisomers

Topoisomer distribution preparation and gel electrophoresis were performed as described previously (*11*) with the following parameters. We prepared different linking numbers by relaxation with calf thymus topoisomerase Ib (Invitrogen) in ethidium bromide concentrations of 0.75 ug/mL, 1.5 ug/mL, 2.25 ug/mL, 3 ug/mL, and 3.75 ug/mL. Each linking number preparation had a total volume of 40 uL and contained 2.7 ug pCR2.1 plasmid derivatives (containing stickleback test sequence, as described in "Plasmid construction" above). They were ethanol precipitated and resuspended in 20 uL of TE pH 8.0 (estimated final concentration ~100 ng/uL). 5 uL of each linking number preparation was mixed to create a 25 uL topoisomer distribution sample, which was heated at 60°C for ~45 min with cap open to reduce volume to ~8 uL. This sample was run on a 13x14 cm 1% agarose gel in a 1.5 mm diameter well in 1x TAE (without ethidium) at 45 V for ~17 h at room temperature. The gel was then washed in 1x TAE with 2 ug/mL chloroquine diphosphate for ~1 h, rotated 90°, and run again in 1x TAE with 2 ug/mL chloroquine diphosphate at 45 V for ~17 h at room temperature. Gels were stained with ethidium and imaged by UV light.

Yeast strains and yeast artificial chromosomes

A list of strains used and generated in this study is provided (table S3). All strains generated in this study were derivatives of the CFY1700 strain, which was generously provided by Catherine Freudenreich. The CFY1700 strain is in a S288c BY4741 background and contains YAC-VS5 (*32*) that has been allowed to break and heal at the T_4G_4 telomeric seed site, so that it can be used to put a new test sequence on the yeast artificial chromosome. Our strains were generated by transforming NotI-linearized pVS20N, pVS20NR, or pVS20N+ori plasmid derivatives (containing stickleback or repeat test

sequence, as described in "Plasmid construction" above) into CFY1700. Transformants were selected on CSM-Ura plates, streaked twice for single colonies on CSM-Leu-Ura plates, and screened for correct integration by Southern blot and PCR.

Yeast artificial chromosome breakage assay

Yeast fragility assays were performed similarly to previous descriptions (*14*). Yeast artificial chromosome strains were grown to single colonies on CSM-Leu-Ura plates. A fluctuation analysis was done by inoculating 10 separate 1 mL liquid cultures of CSM-Leu (+Ura), which were grown at 30°C for 16-18 h to allow breakage to occur. Portions of each culture were plated on CSM (to count total number of cells) and on CSM+FOA-Leu (to count the number of FOA^R cells). We verified that FOA^R resulted from yeast artificial chromosome breakage and complete loss of *URA3* rather than from point mutations (fig. S11). The rate of yeast artificial chromosome breakage is approximated by the rate of FOA^R, which was calculated using a modified method of the median (*34*); a single measurement replicate is the one median value derived from the 10 cultures. For each construct, 2 independent transformant lines were each tested 3 times, in total representing 6 replicates derived from 60 cultures. These median values are plotted in the box plots in Fig. 2, Fig. 3, and fig. S1, and are listed in table S4. Statistical significance calculated using a variety of methods (Wilcoxon Two Sample test for unpaired data, unpaired *t* test, and ANOVA with post-hoc Tukey's Honestly Significant Difference) are listed in full in table S5.

Mammalian mutation assays and LM-PCR

Plasmid DNA (pSP189, pSP189-$(TG)_{41}$, pSP189-$(CA)_{41}$, pSP189-$(TG)_{30}$, or pSP189-$(CA)_{30}$) was transfected into COS-7 cells using GenePORTER according to manufacturer's protocols (GenePORTER, Genlantis Inc., San Diego, CA). After 48 h, plasmids were recovered using a Qiagen Miniprep kit and digested with DpnI to remove unreplicated plasmids. Mutants were identified in MBM7070 bacterial cells by blue-white screening, as previously described (*16*). LM-PCR was carried out 48 h post transfection as previously described (*19, 16*). Briefly, the plasmids were recovered from COS-7 cells using Hirt's method, and the isolated DNA was treated with PolI Klenow Fragment in the presence of dNTPs to blunt the broken ends. After a linker was ligated to the breakpoints, PCR was used to amplify the regions between the specific upstream primer (located 183 bp upstream of the EcoRI site) and the linker. Amplified PCR products were separated by electrophoresis on 1.5% agarose gels.

Stickleback CRISPR oligos

DNA oligos used for stickleback CRISPR mutagenesis are listed below. Uppercase letters in target oligos denote sequence from *Pel*, and target oligo sequences were designed with CHOPCHOP (*35*). Scaffold sequence was previously described (*36*).

Target-Pitx1-129,561.F:

5′- aattaatacgactcactataggAGCCTGATGTGCAGCACACCgttttagagctagaaata

Target-Pitx1-129,601.R:

5′- aattaatacgactcactatagGCACAGTGAAAGGATCCTCCgttttagagctagaaatag

Target-Pitx1-130,342.R:

5′- aattaatacgactcactatagGCTACCTGTTAGCGGCTAGCgttttagagctagaaatag

Target-Pitx1-130,398.R:

5′- aattaatacgactcactataGGCGAGACAGAACCAGAACCgttttagagctagaaatagc

Scaffold:

5′- AAAAGCACCGACTCGGTGCCACTTTTTCAAGTTGATAACGGACTAGCCTTATTTTAACTTGCTA TTTCTAGCTCTAAAAC

Stickleback CRISPR knockout

Guide RNAs for CRISPR/Cas9-mediated mutagenesis were synthesized by *in vitro* transcription. DNA template was created by PCR-mediated extension of a target oligo and HPLC-purified scaffold oligo as follows: 25 uL total of 1 uM target oligo, 1 uM scaffold oligo in Phusion Master Mix (New England Biolabs M0531S), 10 sec 95 °C, 10 sec 60 °C, 10 sec 72 °C, cycled 40 times. PCR product was purified by gel extraction (Qiagen). Template guide DNAs were *in vitro* transcribed to produce guide RNA (gRNA) using the HiScribe T7 High Yield RNA Synthesis Kit (New England Biolabs E2040S) according to manufacturer's directions, treated with DNase, precipitated with LiCl, resuspended in water to 1 ug/uL, and aliquoted at –80°C. Targeted mosaic F_0 sticklebacks were generated by microinjection of freshly fertilized Rabbit Slough fish (*6*). Rabbit Slough is a natural pelvic complete marine population from Alaska that is homozygous for intact *Pel* alleles. The injection mix consisted of 6 uM Cas9-NLS protein (QB3 MacroLab, Berkeley, CA) and 300 ng/uL gRNAs (all four gRNAs transcribed from the target oligos above were pooled; concentration of each single gRNA was 75 ng/uL) in 10 mM Tris-HCl pH 8 with phenol red. Mosaic F_0 individuals were raised to adulthood and crossed to Paxton Lake Benthic fish (a natural population with homozygous *Pel* deletion) to produce $Pel^{WT}/Pel^{PAXB\text{-}Deletion}$ and $Pel^{CRISPR}/Pel^{PAXB\text{-}Deletion}$ F_1 progeny; crosses which did not produce any $Pel^{CRISPR}/Pel^{PAXB\text{-}Deletion}$ progeny (no germline transmission) were discarded. All individuals phenotyped in fig. S5 are siblings from the same cross. *Pel* genotype was determined using nested PCR, with Pitx1-128.592.F and Pitx1-132,268.R as outer primers, and Pitx1-129,112.F and Pitx1-130,662.R as inner primers, and reaction conditions as described for "Plasmid construction" above. Additional mosaic F_0 individuals were crossed to Rabbit Slough or Matadero Creek fish, and Pel^{CRISPR}/Pel^{WT} F_1 progeny DNA was amplified by PCR (nested PCR with Pitx1-128,592.F and Pitx1-132,268.R as outer primers, and Pitx1-129,065.F and Pitx1-131,758.R as inner primers) and sequenced to observe *Pel* CRISPR-induced mutations.

Skeletal preparation

Fish were fixed in 10% neutral buffered formalin for at least 1 week, placed in distilled water (dH2O) for 24 hours, and then placed in 70% ethanol for at least 1 day. Fish were slowly rehydrated to water in a series of at least 1 hour washes (50% ethanol, 25% ethanol, dH2O). Fish were washed twice in 30% saturated sodium borate for 5 minutes and then cleared in trypsin solution (0.25% trypsin in 30% sodium borate) until translucent. Fish were then washed in 2% KOH twice for 5 minutes, and then stained in Alizarin Red solution (0.002% Alizarin Red S powder in 2% KOH) for 24 hours. Fish were bleached in H_2O_2 solution (0.375% KOH, 25% glycerol, and 0.0015% H_2O_2) until pigment was gone. Fish were transferred to 100% glycerol through a series of 0.5% KOH:glycerol solutions (3:1, 1:1, 1:3) and finally stored in 100% glycerol with thymol.

Replication timing materials and library preparation

In brief: Developing stickleback embryos were dissociated and sorted for S-phase and G0/G1-phase cell populations. DNA was extracted from each population and sequenced. In a mixed S-phase population, regions that replicate earlier are at higher copy number (up to 2x) than regions that replicate later. The read depth in S-phase, normalized to the read depth in G-phase, thus represents replication timing. Peaks represent presumptive replication origins or clusters of origins. Steep slopes indicate primarily unidirectional replication. In detail: Stickleback clutches of ~50-100 embryos from 9 different populations (BDGB, RABS, JADE, BOOT, BEPA, KFSY, LSHP, ORPH, BOUL) of wild-caught fish were grown for ~10 days at 16°C until they reached Swarup developmental stages 24-26 (*37*). Each clutch was then dissociated into single cells using the Worthington Papain Dissociation System according to the manufacturer's instructions with ~40 minutes incubation in papain solution. Single cells were resuspended in 3 mL cold PBS with 1% FBS, and 7 mL of cold 100% ethanol was added slowly while swirling to fix the cells. Fixed cells were stored at –20°C. Ethanol fixed cells were spun down and resuspended in 50 ug/mL propidium iodide, 250 ug/mL RNaseA, in PBS with 1% FBS to stain for DNA content. Cells were filtered through a 70 um nylon cell strainer before sorting on a BD FACS Aria by propidium iodide. Although traditional replication timing strategies collect an early S-phase and a late S-phase population (*38*), we collected just one mid S-phase population, due to the limiting number of cells undergoing DNA replication in developing embryos. Sorting continued until ~2,000,000 G0/G1-phase cells and ~250,000 S-phase cells were collected. Another 200 ug of glycogen was added to the collected cells along with ~3-5 volumes of 100% ethanol, and then cells were precipitated by centrifugation at 300 *g* for 5 min. The pellet was resuspended in 600 uL of 10 mM Tris pH 8.0, 100 mM NaCl, 10 mM EDTA, 0.5% SDS, and 333 ng/uL Proteinase K, and incubated overnight at 55°C. Genomic DNA was isolated by phenol:chloroform extraction with 1-2 chloroform washes and ethanol precipitation. Sequencing libraries were prepared using the Nextera DNA Library Preparation Kit (Illumina) according to the manufacturer's instructions, except that we increased Tagmentation time to 10 minutes from 5 minutes, and we used 50 uL of AMPure XP beads instead of 30 uL for PCR cleanup. G0/G1-phase DNA and S-phase DNA were separately barcoded during library preparation and sequenced together in one lane (per population) on the Illumina HiSeq 2000 platform with single-end 101 bp runs.

Replication timing analysis

Reads were adapter trimmed and aligned to the reference stickleback genome (version gasAcu1) using Burrows-Wheeler Aligner's Smith-Waterman Alignment (*39*). Read depth was extracted for each base in the genome, and bases mapping more than 3 standard deviations over the mean read depth were discarded from subsequent analysis. Each individual G0/G1-phase and S-phase library read depth was normalized to its own mean depth, then all 9 populations' G0/G1-phase normalized depths were combined into a single G0/G1-phase dataset, and all 9 populations' S-phase normalized depths were combined into a single S-phase dataset. The S/G depth ratio was calculated using the combined normalized read depth from these datasets over a 50 kb sliding window with 25 kb step. Slopes were calculated using 5 windows upstream and 5 windows downstream. Raw sequencing data and processed S/G read depth ratio data have been deposited at the Gene Expression Omnibus, accession GSE121537.

TG-repeat identification

TG-repeats in the genome were identified using Basic Local Alignment Search Tool (BLAST version 2.2.31). A $(TG)_{100}$ sequence (i.e. 200 nucleotides of "TGTGTGTG…") were aligned by BLAST against the stickleback (version gasAcu1) or human (version hg19) reference genomes using default settings without repeat masking. Since the query sequence is repetitive, raw BLAST results will return multiple matches for a single stretch of genomic repeats (e.g. a single $(TG)_{33}$ repeat at chrI:28680-28745 will match query positions 1-66, 2-67, 3-68, etc., and also query positions 1-60 will match chrI:28680-28739, chrI:28681-28740, chI:28682-28741, etc.). Thus, results with overlapping chromosome positions were collapsed into a single entry. Because overlapping results were collapsed, repeat stretches both shorter or longer than $(TG)_{100}$ were identified, and the distribution in identified repeat lengths did not change if $(TG)_{50}$ or $(TG)_{200}$ query sequences were used instead of a $(TG)_{100}$ query sequence.

Human double-strand break analysis

Human aphidicolin sensitive double-strand break sites (*40*) and human replication timing data (*41*) were obtained from previous studies and from ENCODE (GEO GSM923449). The top 5,000 aphidicolin sensitive 48-kb genomic windows were used in the enrichment analysis. Replication direction slope for each TG-repeat was calculated using 5 replication timing windows upstream and 5 windows downstream of the location of the TG-repeat. TG-repeats were then split into 3 quantiles for replication direction slope and 3 quantiles for TG-repeat length. Each of these 9 groups contained ~5,000 TG-repeat stretches. Enrichment and statistical significance were calculated as described previously (*42*). Briefly, the null overlap distribution was calculated by holding the locations of TG-repeats in each class constant (~5,000 locations) and reassigning the aphidicolin sensitive windows (5,000 locations) to every possible location in the genome. Enrichment was calculated by dividing the real number of overlaps between TG-repeats and aphidicolin sensitive windows by the average number of overlaps in the null distribution. P values represent the probability of the real number of overlaps or more occurring by chance in the null distribution. Similar results are obtained if the null distribution is calculated by holding the locations of aphidicolin sensitive windows constant and reassigning the locations of TG-repeats.

Population genetics modeling

The number of potential de novo mutations at a specific locus in the genome was calculated for a range of possible mutation rates (μ) and stickleback census population sizes (N). The modeled mutation rates (10^{-9}-10^{-2}) span spontaneous mutation rates for a range of mechanisms (*43, 44*). The modeled population sizes (10^{2}-10^{9}) generously cover estimates for a range of stickleback freshwater habitats (10^{3}-10^{6}), from small ponds to large lakes (*45*). The total number of generations (G) was set to 10^{4}, as expected for post-glacial populations breeding once every one to two years (*45, 46*). The total number of potential mutations arising at a particular locus (Θ) was calculated by multiplying the population mutation rate ($\theta = 2N\mu$) by the number of generations, as follows:

(*Equation 1.*) $$\Theta = 2N\mu G$$

The probability that any given mutation will eventually fix at any time in the future (π) was calculated using Kimura's general diffusion equation (*21*) with additivity (dominance coefficient $h = 0.5$), since *Pel* mutations are semidominant based on previously published empirical results (*47*):

(*Equation 2.*) $$\pi = (1 - e^{-sN_e/N}) / (1 - e^{-2sN_e})$$

where N_e is the effective population size, and s is the selection coefficient (relative fitness of the homozygote). Effective population sizes are typically 10% or less of census population sizes (*48*), and have been previously estimated as 3-6% in some stickleback populations (*49, 50*). A conservative estimate of $N_e/N = 0.1$ was used for Fig. 4D and fig. S7. Results using a wide range of N_e/N (from 0.01 to 1) are shown in fig. S6, with the general result that successful adaptation probability being highly dependent on mutation rate holding true even at the extreme case of $N_e = N$, albeit less so at the largest population sizes and highest selection coefficients. Positive selection on a new pelvic reduction allele was modeled in the range of $s = 0.1$ to $s = 0.001$. Previous studies have measured strong selection on some stickleback armor traits, including $s \geq 0.1$ for the *Eda* major plates locus (*51*). For comparison, pelvic reduction requires ~2,000 generations to appear and reach high frequency in the fossil record (*52*), corresponding to time-averaged s on the order of ~0.01 or less. The neutral scenario (genetic drift) was modeled with $\pi = 1/2N$ (which is also the limit of π as s approaches 0 in Equation 2).

The probability that at least one mutant allele will arise within G generations that will successfully fix in the future, as shown in Fig. 4D, fig. S6, and fig. S7A, is:

(*Equation 3.*) $$1 - (1 - \pi)^{\Theta}$$

However, for neutral and small s, the average time to fixation for an eventually successful allele will exceed the total number of generations under consideration (*53*) (fig. S7B). We therefore integrated the known probability density function of times to fixation for a neutral allele (*54*), to calculate the cumulative distribution function of times to fixation for a neutral allele. The probability that a neutral allele destined for fixation will fix in $t \leq G$ generations is then:

(*Equation 4.*) $$Y(G) = \int_{t=0}^{G} \sum_{i=1}^{\infty} (2i+1)(-1)^{i+1} \lambda_i e^{-\lambda_i t} dt$$

where $\lambda_i = i(i+1)/4N_e$. Although the distribution of times to fixation is classically described for neutral alleles (*54*), it is not for selectively advantageous alleles ($s > 0$). However, for the $G = 10{,}000$ generations considered here, the average time to fixation for $s = 0$ and $s = 0.001$ is similar (fig. S7B). We therefore also used Equation 4 for the $s = 0.001$ condition. For larger s ($s = 0.01$ and $s = 0.1$), the average time to fixation is small enough compared to $G = 10{,}000$ generations (fig. S7B) that we did not model this additional constraint.

The adjusted probability that at least one mutant allele will arise within G generations and also have time to fix within G generations, as shown in fig. S7C, is therefore:

(*Equation 5.*) $$1 - (1 - \pi\, Y(G))^{\Theta}$$

This calculation does not account for the reality that, e.g., a mutant arising at generation 6,000 will only have 4,000 generations left (and not the full 10,000 generations) until 10,000 generations have passed from the initial colonization. The final probabilities are therefore a slight overestimate but do not alter the interpretation of the results.

Mutational mechanisms in human evolution

To examine mutational mechanisms contributing to likely adaptive traits in humans, we focused on a time frame of migration of modern humans out of Africa to new regions around the world (occurring over roughly 60,000 years or 3,000 generations, with estimated effective population size of 10^4, conditions similar to evolutionary parameters in sticklebacks (*24, 55*)). We identified 94 examples of human traits likely to be adaptive and having a known molecular basis (table S1), based on an updated review of previously known loci of evolution (*56*). We further classified molecular changes based on whether they affect protein coding or non-coding regions, and whether they had likely arisen by low mutation rate mechanisms (small indels and single nucleotide changes at non-CpG sites) or high mutation rate mechanisms (homopolymer slippage, large indels, and C to T single nucleotide changes at CpG sites, which are known to occur at rates ~10-18 times higher than that of other substitutions (*43, 44*)).

Figure S1

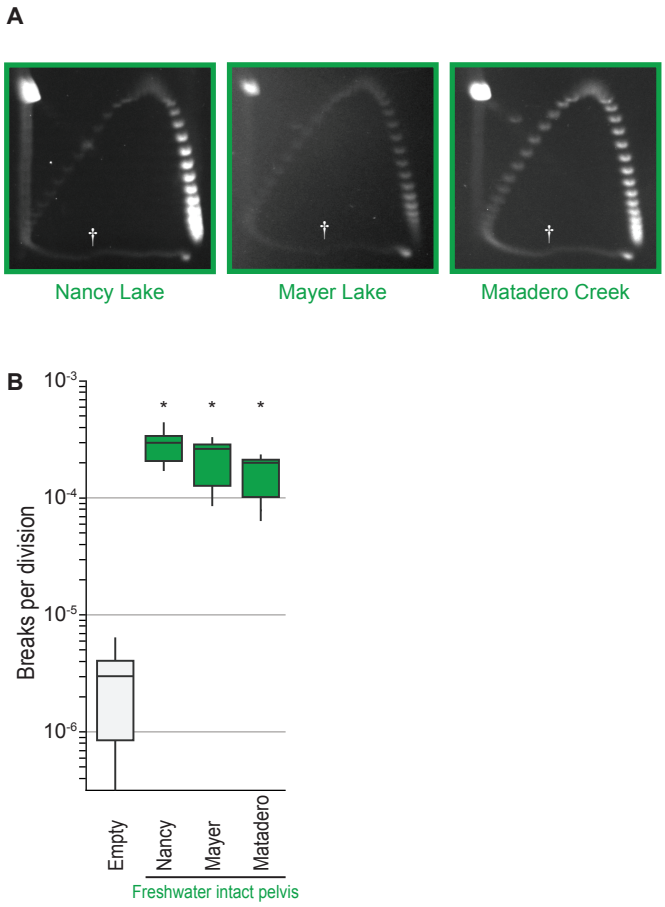
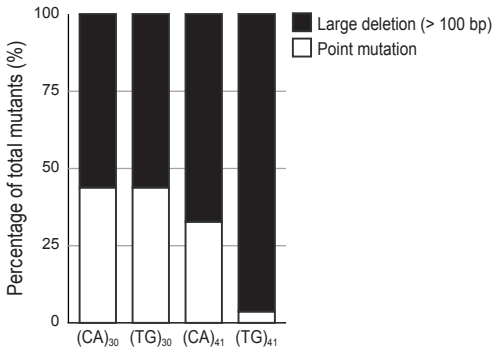

A
Nancy Lake
Mayer Lake
Matadero Creek
B
Breaks per division
10^{-3}
10^{-4}
10^{-5}
10^{-6}
*
*
*
Empty
Nancy
Mayer
Matadero
Freshwater intact pelvis

Figure S1. *Pel* DNA sequences from freshwater pelvic-complete fish retain alternative structure and high breakage rates.

(A) *Pel* sequences from three different freshwater populations without pelvic reduction. Daggers, mobility shifts. **(B)** Yeast artificial chromosome breakage rates from the same populations, plotted as in Fig. 2. *$p < 0.01$ (table S5). Although pelvic reduction is only seen in freshwater stickleback populations, not all freshwater populations have pelvic reduction. Environmental conditions that favor pelvic reduction are rare (*5, 57*). Low mutation rates may preclude the fixation of a trait when it is selectively advantageous (Fig. 4D, fig. S6), but high mutation rates alone do not dictate the path of evolution when a trait is disadvantageous. That elevated mutation rates are important in stickleback evolution is therefore compatible with and does not contradict classic descriptions of mutation-selection balance or the idea that environmental conditions also play a key role in evolutionary outcomes.

Figure S2

A



B

Deletion junctions	Microhomology	Other mutations
Unknown - 147		
-11 - 217		
6 - 260	C	
6 - 260	C	
6 - 261	C	
7 - 151	GACTTCG	
21 - 213	G	
29 - 160	GGT	TGT to GGT point mutation at deletion junction
30 - 161	GT	
38 - 149	T	
38 - 234		
43 - 247		
43 - 247		
49 - 256	CGG	Insertion of sequence from 250-277 at deletion junction
49 - 256	CGG	Insertion of sequence from 250-277 at deletion junction
52 - 247	GACCTGAC	Insertion of sequence from 4093-4125 at deletion junction
56 - 102		
94 - 239	TGAGCTCGA	
162 - 851		
Unknown - 99		Missing 26 TG repeats
Unknown - 102		Missing 23 TG repeats
Unknown - 122		Missing 23 TG repeats
Unknown - 123		Missing 23 TG repeats

C

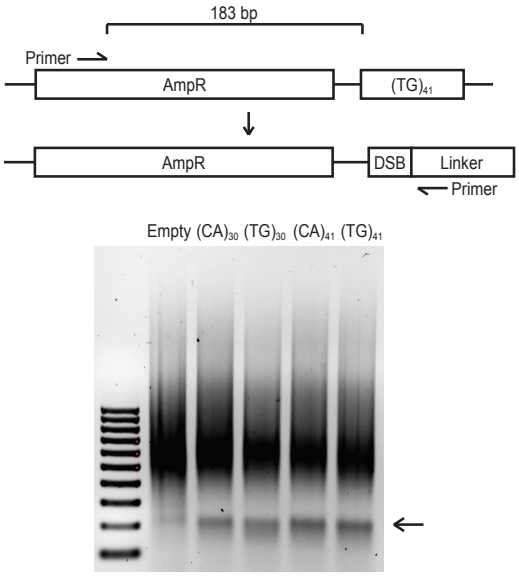


Figure S2. TG-repeat-induced mutations and double-strand breaks in mammalian cells.

(A) Mutation spectrum of sequenced mutants. This assay is sensitive to most types of mutations, including point mutations and deletions. Nearly all of the mutations induced by $(TG)_{41}$ are >100 bp deletions. **(B)** Detailed table of all sequenced $(TG)_{41}$-induced deletions in mammalian cells, a subset of which (the clean deletions) are shown in Fig. 3F. Unknown indicates junctions that extended near the primed region. Repeated rows indicate same deletion recovered multiple times. Coordinates are defined such that the $(TG)_{41}$ repeat is from positions 7-88. **(C)** The junctions of the deletions suggested that the mutagenic events might occur during repair of breaks induced by the repeat sequences. We further mapped the breaks using ligation-mediated PCR (LM-PCR) on plasmids recovered from mammalian cells 48 hours after transfection. Schematic shows the location of the upstream PCR primer (located in the mutation shuttle vector) and a downstream PCR primer (located in a linker sequence that is added by ligation at the position of breaks). Amplified products were separated on a 1.5% agarose gel. All 4 repeats lead to the production of PCR products of ~210 bp, suggesting the formation of breaks ~190-bp downstream of the specific primer, which is near the dinucleotide repeat sequences.

Figure S3

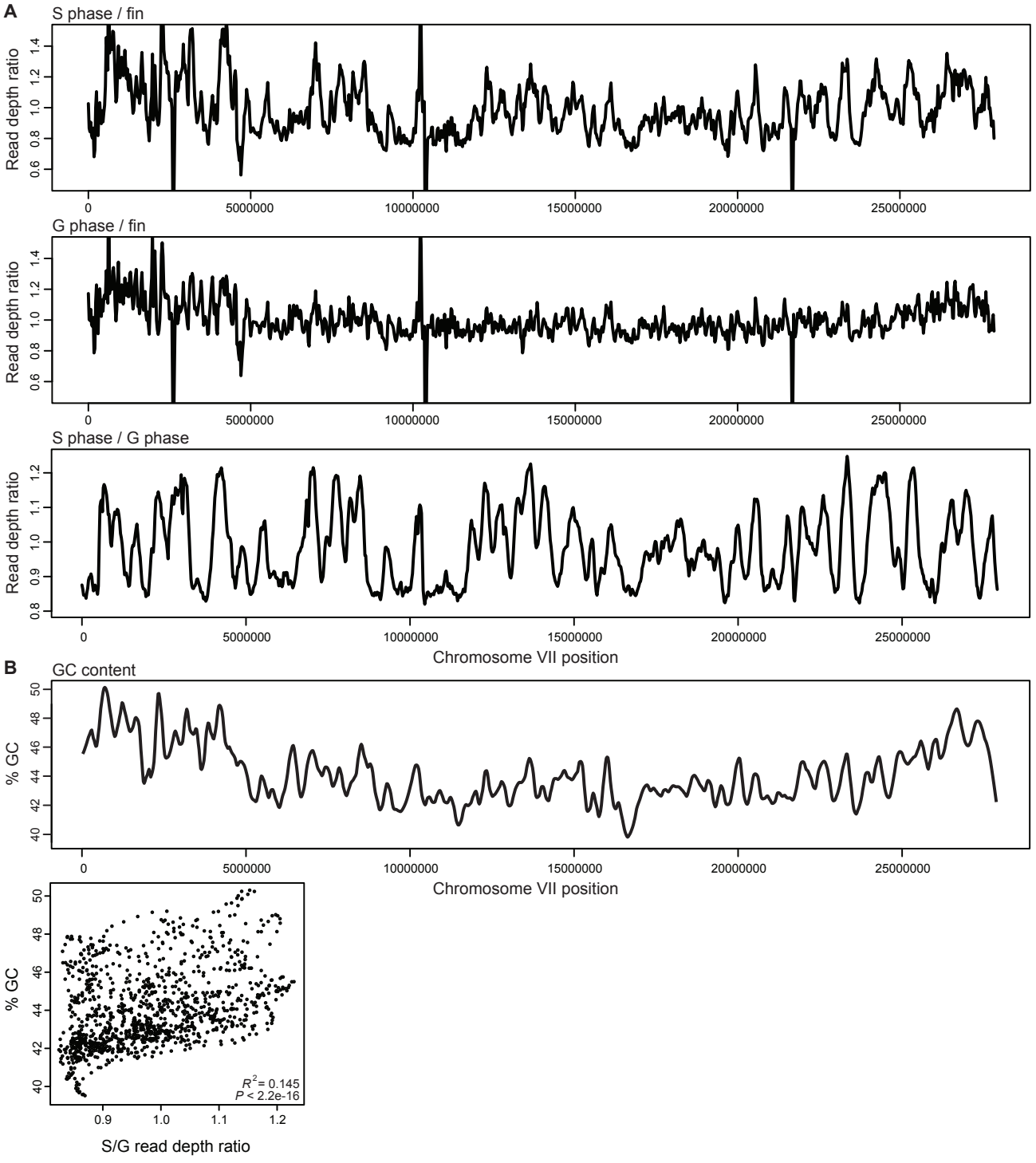

A
S phase / fin
Read depth ratio
G phase / fin
Read depth ratio
S phase / G phase
Read depth ratio
Chromosome VII position
B
GC content
% GC
Chromosome VII position
% GC
R^2= 0.145
P < 2.2e-16
S/G read depth ratio

Figure S3. Stickleback replication timing profile validation.

(A) Replication timing signal is driven by S-phase read depth, as theoretically expected (Fig. 4A). Chromosome VII S-phase read depth (top panel) or G-phase read depth (middle panel) was normalized to adult non-dividing fin tissue read depth. The full chromosome VII S/G profile from Fig. 4 is provided (bottom panel) for reference. The S/fin profile is similar to that of S/G, whereas the G/fin profile is noisy. **(B)** Top panel, GC content profile along chromosome VII. Bottom panel, chromosome VII S/G read depth ratio vs. GC content. Earlier replicating regions (higher S/G ratio) tend to have higher GC content, as reported previously (*58*).

Figure S4

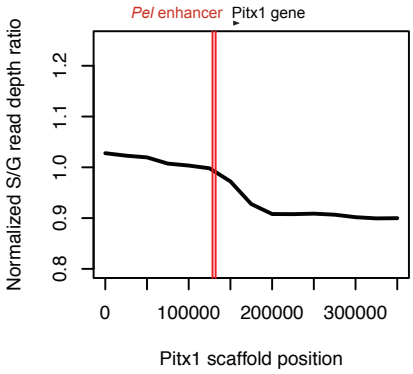
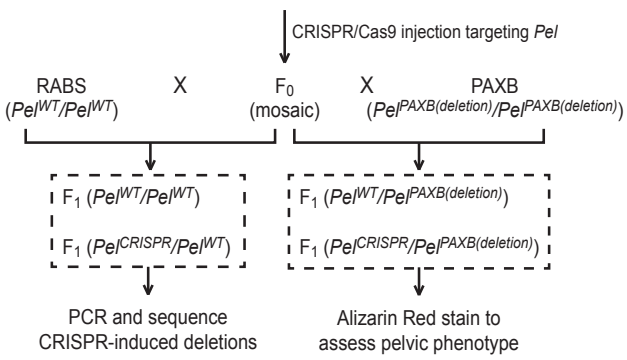

Pel enhancer
Pitx1 gene
Normalized S/G read depth ratio
0.8
0.9
1.0
1.1
1.2
0
100000
200000
300000
Pitx1 scaffold position

Figure S4. Close up of the end of chromosome VII containing the *Pel* locus.

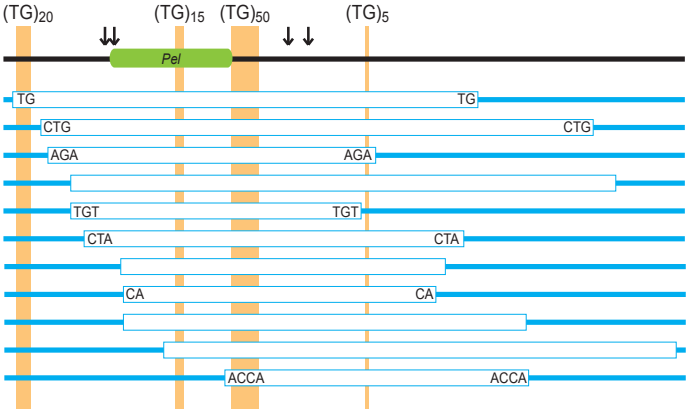
Scaffold position 0 starts immediately after assembly gap shown in Fig. 4B. Red lines denote beginning and end of *Pel* region tested in Fig. 1, Fig. 2, and diagrammed in Fig. 3A. Black arrowhead denotes position of *Pitx1* gene, which transcribes to the right. The position of the *Pel* sequence to the right of a replication timing peak indicates that the TG repeat-rich strand of the *Pel* enhancer would serve as the template for lagging strand DNA synthesis, which is the fragile orientation based on fragility assays (Fig. 2, Fig. 4C).

Figure S5

A



B *Pel* region CRISPR-induced deletions

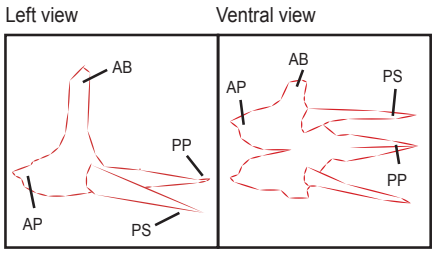


C

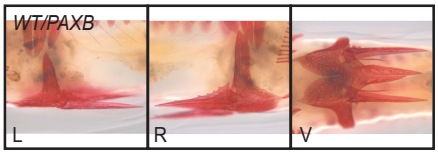
Pel genotype	Full pelvis	Reduced pelvis
WT / PAXB	37	0
CRISPR / PAXB	7	7

$p < 0.0001$ Fisher's exact test

Full pelvis diagram



Representative full pelvis staining:



Reduced pelvis staining:

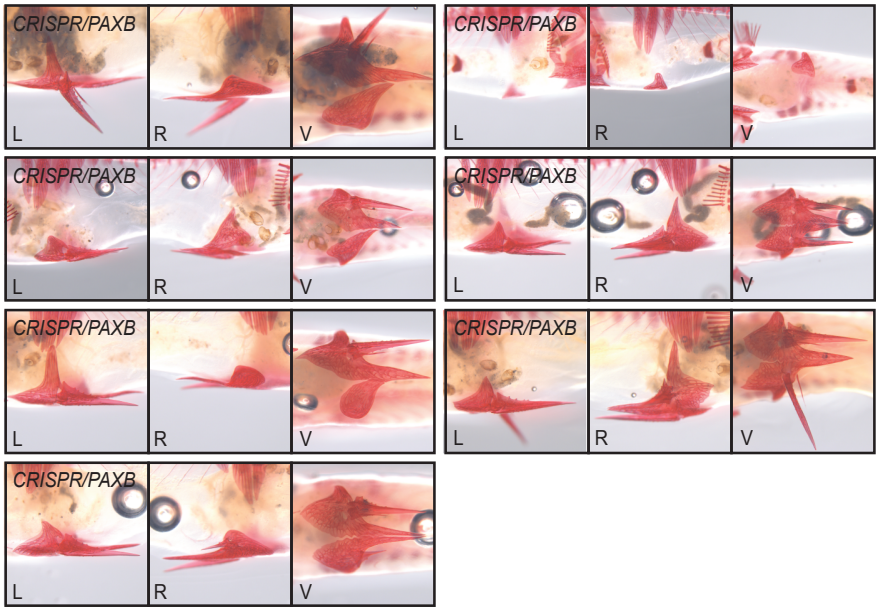


Figure S5. CRISPR/Cas9 targeting of *Pel* sequence causes deletions and pelvic reduction in stickleback crosses.

(A) Diagram of cross strategy for detecting deletions and pelvic phenotypes. Fertilized eggs from the RABS marine population (wild type at *Pel*) were injected with Cas9 and guide RNAs flanking the *Pel* enhancer region. Mosaic F_0 founders were crossed to RABS to transmit deletion alleles and characterize typical lesion breakpoints (deletion alleles are smaller than wild type alleles and easier to amplify in the presence of the RABS allele). An additional F_0 founder was crossed to the PAXB freshwater population to test for pelvic reduction phenotypes (PAXB fish carry a natural deletion of *Pel*, making it possible to score recessive pelvic phenotypes in F_1 hybrids without additional generations of breeding). **(B)** To-scale map of CRISPR/Cas9-induced deletions transmitted to F_1 hybrids in crosses with RABS. Green box, *Pel* sequence previously shown to drive pelvic expression (*6*). Light brown shading, location of TG-repeat sequences. Arrows, location of guide RNAs flanking *Pel* region. White boxes, DNA deletions. Blue lines, DNA remaining. Letters indicate microhomologies present at deletion junctions. **(C)** Genotypes and phenotypes in cross with PAXB. All individuals in table are siblings from the same cross. Fish that inherit a PAXB allele and a transmitted wildtype allele all develop a full pelvis. In contrast, fish that inherit a PAXB allele and a transmitted CRISPR-deletion allele show various forms of pelvic reduction. Typical structures of a fully developed stickleback pelvis are labeled in line diagrams in left and ventral views. AB, ascending branch; AP, anterior process; PP, posterior process; PS, pelvic spine. Alizarin Red stained skeletal structures are shown for a representative Pel^{WT}/Pel^{PAXB} animal with a complete pelvis and all seven Pel^{CRISPR}/Pel^{PAXB} animals with various forms of pelvic reduction. Panel views are L, left; R, right; and V, ventral for each individual.

Figure S6

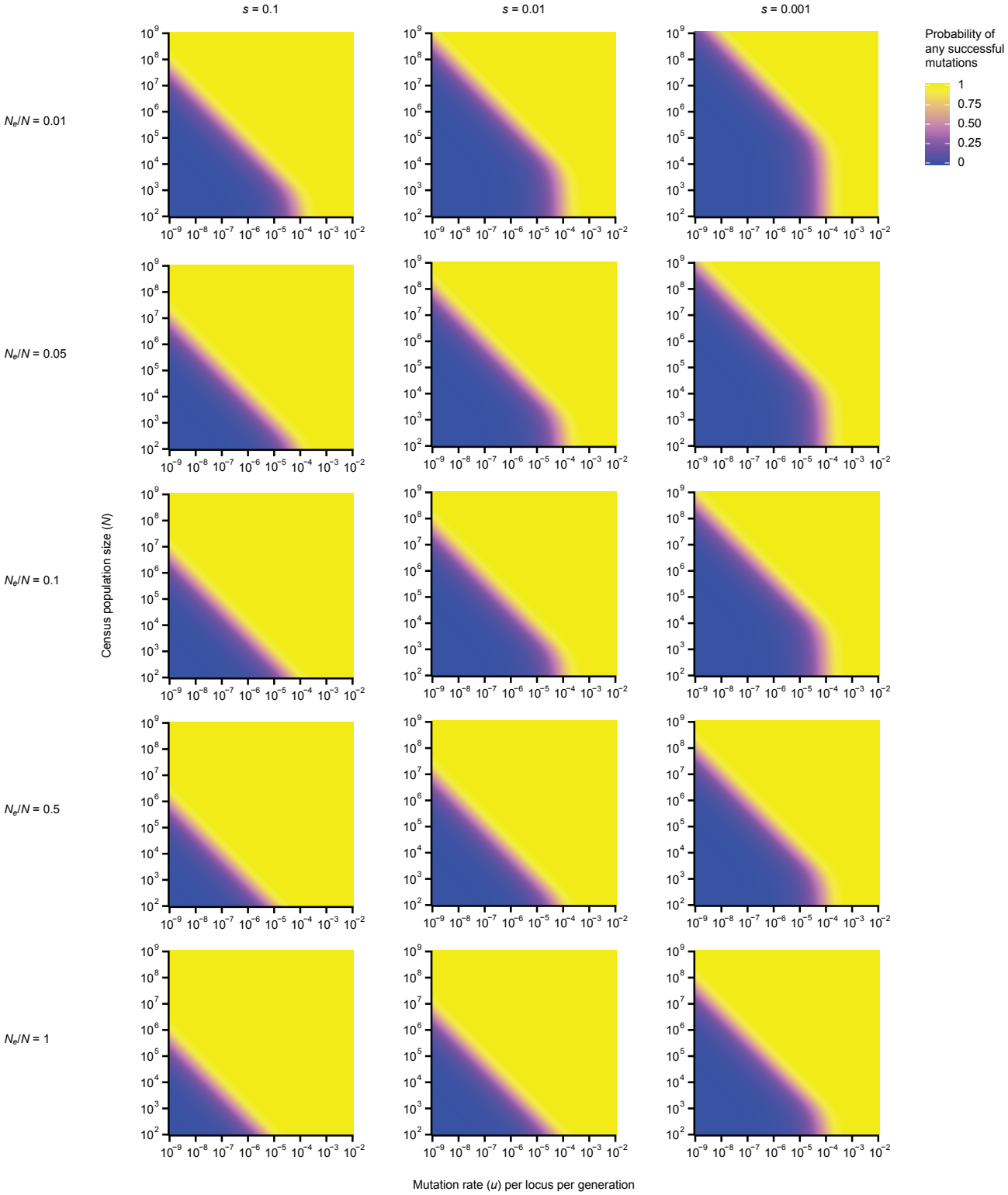

s = 0.1
s = 0.01
s = 0.001
Probability of any successful mutations
1
0.75
0.50
0.25
0
Ne/N = 0.01
Ne/N = 0.05
Ne/N = 0.1
Ne/N = 0.5
Ne/N = 1
Census population size (N)
Mutation rate (u) per locus per generation

Figure S6. Modeling results over a wider range of parameters.

The probability of at least one mutation arising at a particular locus within 10,000 generations and successfully fixing at any time in the future was calculated as in Fig. 4D (also see Methods Equation 3). Results shown here are across a wider range of selection coefficients (s), population sizes (N), and effective population size ratios (N_e/N) than shown in Fig. 4D.

Figure S7

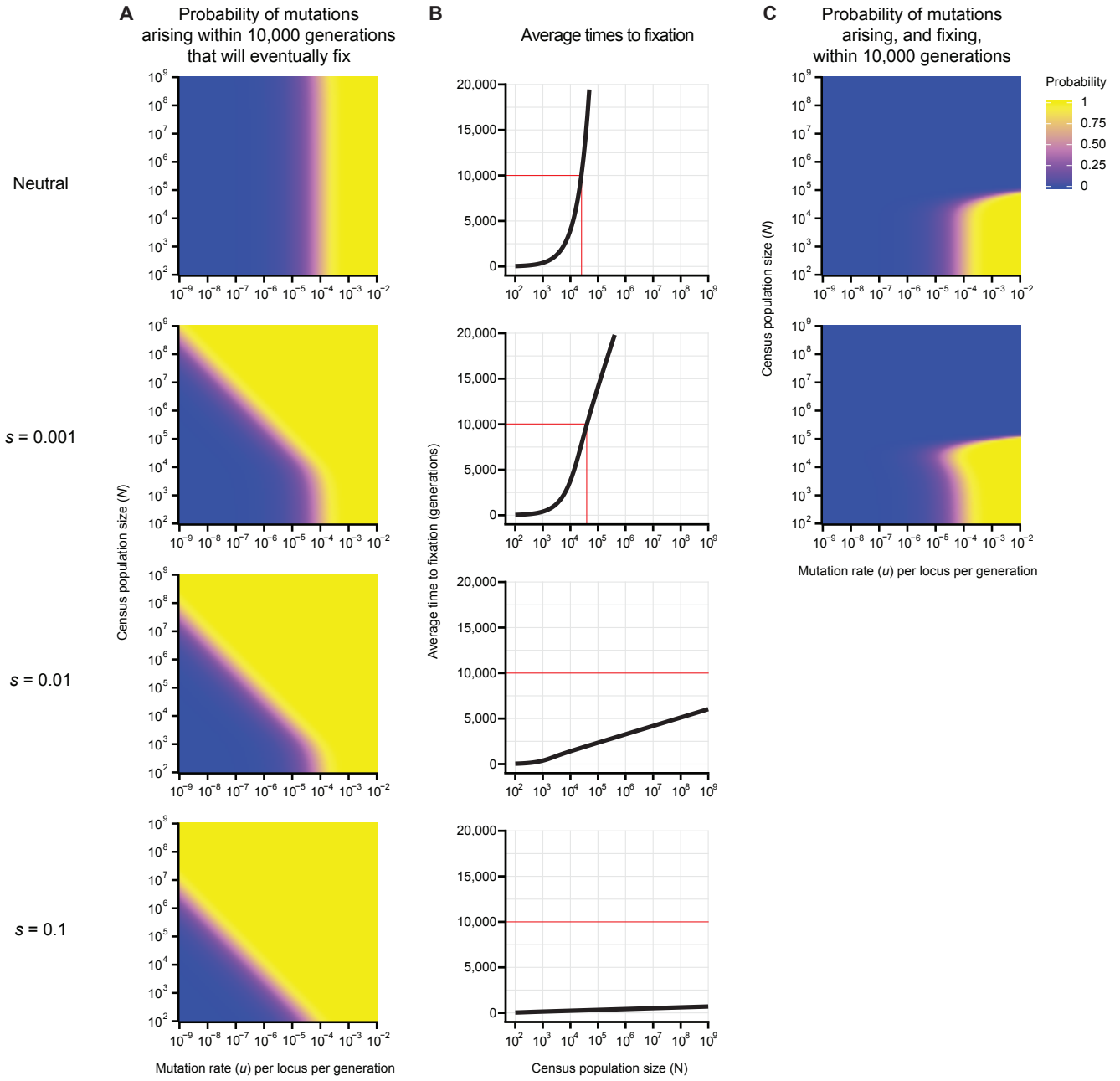

A
Probability of mutations arising within 10,000 generations that will eventually fix
B
Average times to fixation
C
Probability of mutations arising, and fixing, within 10,000 generations
Probability
1
0.75
0.50
0.25
0
Neutral
s = 0.001
s = 0.01
s = 0.1
Census population size (N)
Average time to fixation (generations)
Mutation rate (u) per locus per generation
Census population size (N)

Figure S7. Probability of successful de novo mutation depends on mutation rate.

(A) The occurrence and fixation of different types of spontaneous mutations was modeled in freshwater sticklebacks using G=10,000, N_e/N=0.1, and a range of mutation rates, population sizes, and selection coefficients (also see Methods Equation 3). **(B)** The average time to fixation for an eventually successful allele will exceed 10,000 generations for neutral and small s, when the population size begins to exceed ~10^4-10^5. **(C)** The adjusted probability that at least one mutant allele will arise within G=10,000 generations and also have time to fix within G=10,000 generations (also see Methods Equation 5). Because mutations occurring at the highest fragile site frequencies can fix under both neutral and adaptive scenarios, these results do not provide new information about whether pelvic reduction is adaptive in sticklebacks. However, an adaptive model is supported by several other types of independent ecological and molecular data, including: association of pelvic-reduction with particular ecological conditions in between-lake comparisons (*5*); consistent association of the trait with particular environments within lakes, maintained even in the face of gene flow (*4, 59*); consistent molecular signatures of positive selection surrounding *Pel* deletions in multiple pelvic-reduced populations (*6*); and trait variation that exceeds molecular signatures of neutral drift within populations (*7*). Thus, repeated use of *Pitx1* for pelvic evolution in sticklebacks most likely reflects a combination of both elevated mutation rates making de novo variants available, and ecological conditions providing a selective advantage for pelvic-reduction alleles in particular environments.

Figure S8

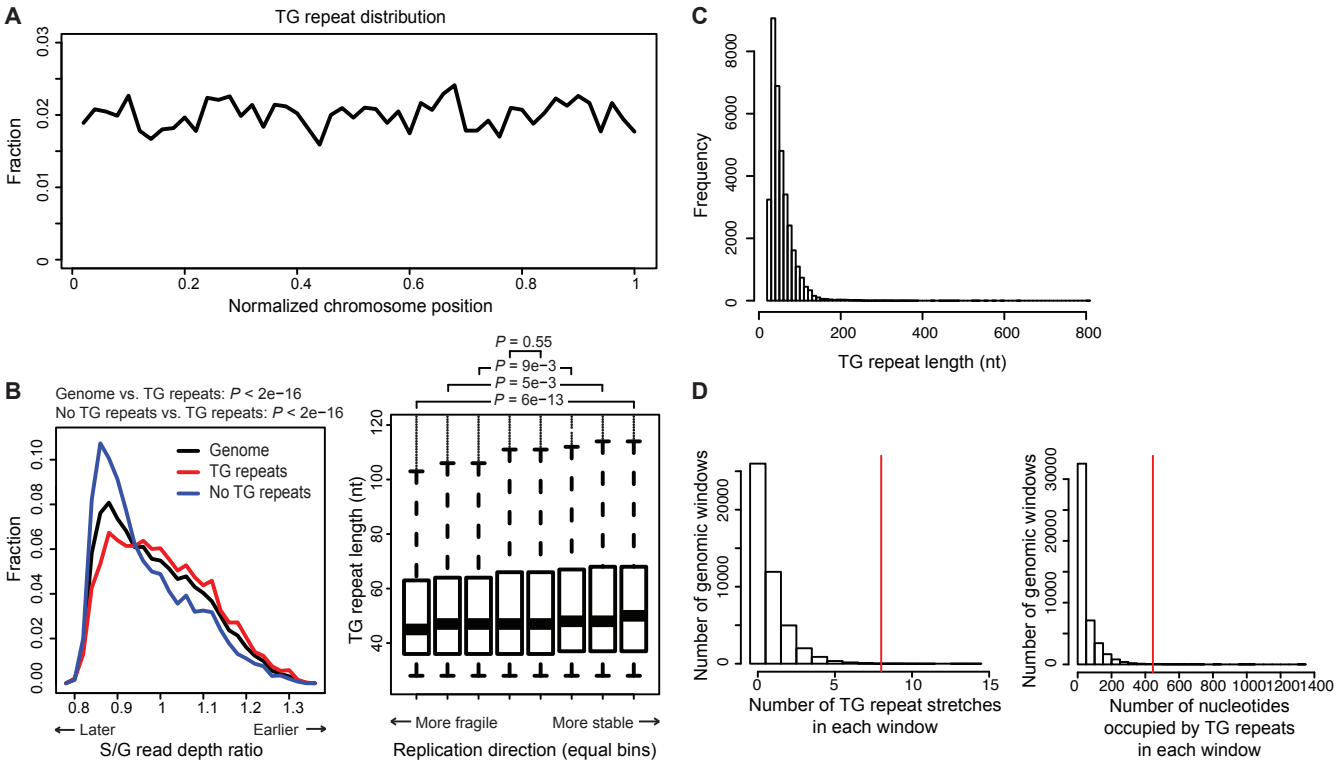

A
TG repeat distribution
Fraction
Normalized chromosome position
B
Genome vs. TG repeats: P < 2e−16
No TG repeats vs. TG repeats: P < 2e−16
Genome
TG repeats
No TG repeats
Fraction
← Later
Earlier →
S/G read depth ratio
P = 0.55
P = 9e−3
P = 5e−3
P = 6e−13
TG repeat length (nt)
← More fragile
More stable →
Replication direction (equal bins)
C
Frequency
TG repeat length (nt)
D
Number of genomic windows
Number of TG repeat stretches in each window
Number of genomic windows
Number of nucleotides occupied by TG repeats in each window

Figure S8. TG-repeats in the stickleback genome.

(A) Distribution of the ~34,000 TG-repeat stretches identified in the stickleback genome along a meta-chromosome, consisting of every chromosome normalized to length 1. **(B)** Left panel, distribution of S/G read depth ratios in 50 kb windows, for either: all windows, only windows that contain TG-repeats, or only windows that do not contain TG-repeats. Regions with TG-repeats are depleted in later replicating parts of the genome. Right panel, distribution of lengths for TG-repeats separated into 8 equal bins. Bins are ordered by replication timing slope, from primarily unidirectional replication in the fragile orientation (most negative slope) to neutral (slope near 0) to primarily unidirectional in the stable orientation (most positive slope). TG-repeats are significantly shorter in genomic regions where they are predicted to be in the fragile replication orientation. P values for both panels were calculated using Wilcoxon Two Sample Test. These small but significant biases could result from either neutral sequence evolution (fragile TG-repeats create DNA breaks, which delete or shorten the repeat) or purifying selection (fragile TG-repeats near essential genes would be disadvantageous) and suggest that TG-repeat-induced fragility has shaped the stickleback genomic landscape. **(C)** Distribution of TG-repeat lengths in the stickleback genome. **(D)** Distribution of the number of TG-repeat stretches and number of total nucleotides occupied by TG-repeats in 10 kb windows. Red line indicates the genomic window containing the *Pel* locus, which contains 4 TG-repeat stretches in the ~3 kb window tested in Fig. 1, Fig. 2, and fig. S1, plus 4 additional stretches in the rest of the 10 kb window. Although the *Pel* region is an outlier in its number of TG-repeats, the breakage assay results show that a single TG-repeat stretch is enough to confer fragility (Fig. 3). Thus, many other locations in the genome harboring TG-repeats in the fragile orientation may also represent sites with elevated mutation rates.

Figure S9

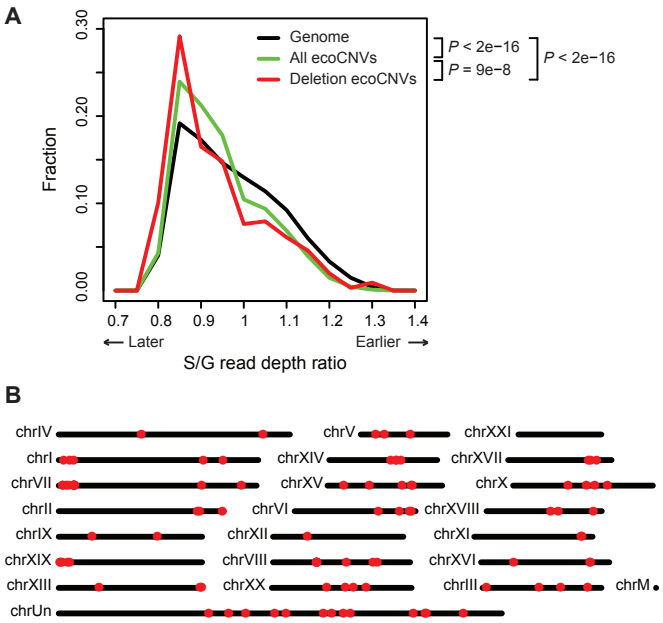

A
Genome
All ecoCNVs
Deletion ecoCNVs
P < 2e−16
P = 9e−8
P < 2e−16
Fraction
0.00
0.10
0.20
0.30
0.7
0.8
0.9
1
1.1
1.2
1.3
1.4
← Later
Earlier →
S/G read depth ratio
B
chrIV
chrV
chrXXI
chrI
chrXIV
chrXVII
chrVII
chrXV
chrX
chrII
chrVI
chrXVIII
chrIX
chrXII
chrXI
chrXIX
chrVIII
chrXVI
chrXIII
chrXX
chrIII
chrM
chrUn

Figure S9. Ecotypic copy number variation in the stickleback genome.

(A) Distribution of S/G read depth ratios for either: the whole genome, the previously identified ~6,500 recurrent sites of copy number variation (CNV) that are consistently different between marine and freshwater ecotypes (ecoCNVs) (*31*), or only ecoCNVs that are deletions. P values were calculated using Wilcoxon Two Sample Test. These ecology/habitat-associated CNVs, especially deletions, were significantly enriched in late replicating regions of the stickleback genome, a trend also noted in humans (*41*). **(B)** Of 659 regions present in marine populations but recurrently deleted in freshwater populations, 98 (14.9%) were near (≤ 1 kb) a TG-repeat (table S1). Genomic locations of these 98 deletions are shown.

Figure S10

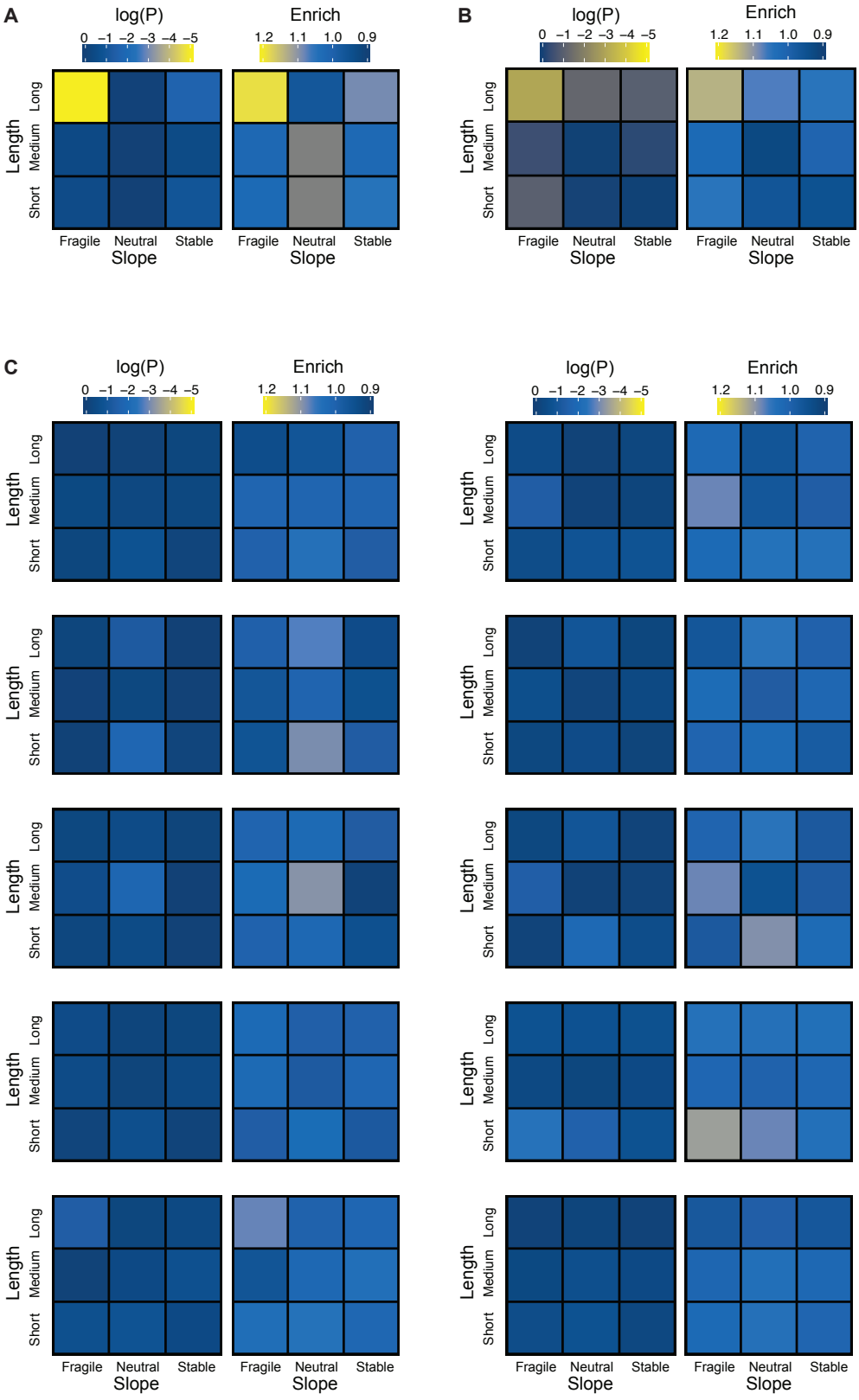

A
log(P)
0 −1 −2 −3 −4 −5
Enrich
1.2 1.1 1.0 0.9
Length
Long
Medium
Short
Fragile
Neutral
Stable
Slope
B
C

Figure S10. Association of human aphidicolin-sensitive DNA double-strand break sites with classes of TG-repeats.

(A) Replication timing data (*41*) was used to determine the direction of replication through ~45,000 human TG-repeats. y-axis, TG-repeats length quantiles. x-axis, replication timing slope quantiles. Left panel, p values. Right panel, enrichment near DNA breaks of long TG-repeats in the fragile orientation. Long TG-repeats in the fragile orientation are slightly but significantly enriched around sites of DNA breakage reported in human cells following exposure to the DNA replication inhibitor aphidicolin. The enrichment is small because TG-repeats are not the only mechanism creating DNA breaks. No enrichment was observed with short repeats or with repeats in stable orientations. TG-repeats may thus also contribute to DNA breakage sites in humans. **(B)** The same analysis as in (A), except using a replication timing dataset from HeLa-S3 cells (GEO GSM923449). HeLa-S3 cells are diverged from HeLa cells in both morphology (suspension instead of adherent) and genotype (both cell lines have unstable karyotypes). The association between long TG-repeats in the fragile orientation and DNA breaks is still significant but less so with HeLa-S3 data, and the p value color scale is adjusted to better visualize the distinction. **(C)** The same analysis as in (A) repeated 10 times, except each time using 5,000 random genomic sites instead of the top 5,000 aphidicolin-sensitive genomic sites. Random genomic windows do not show any significant enrichment with any TG-repeat types.

Figure S11

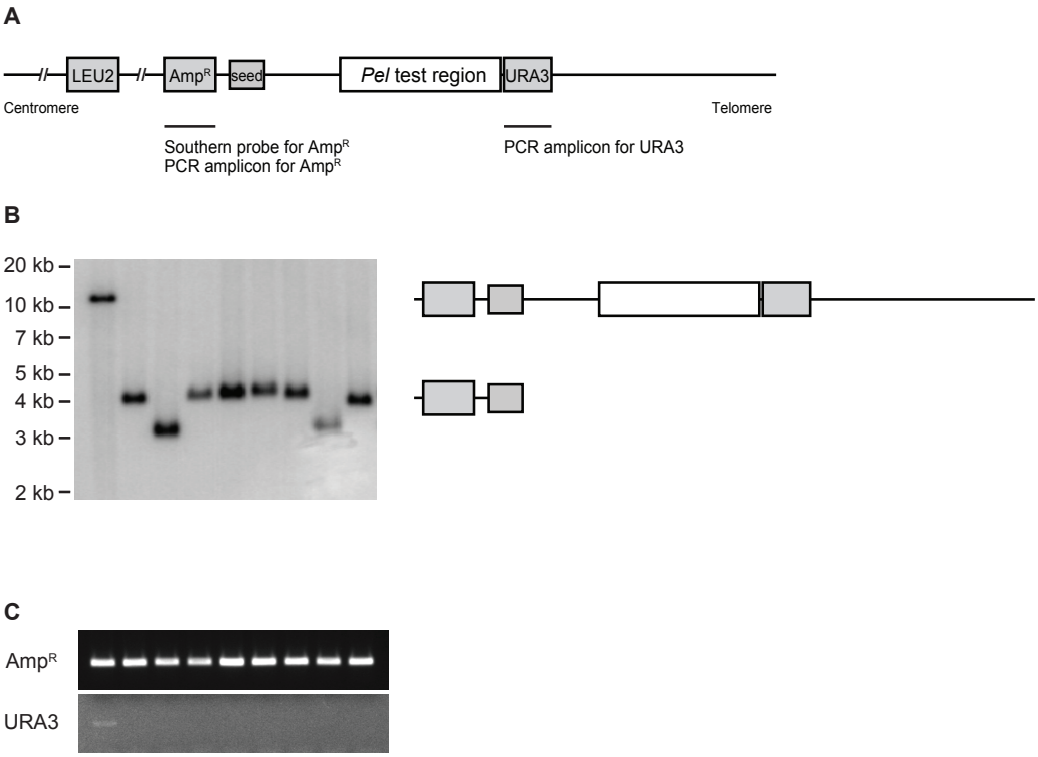

A
LEU2
Amp^R
seed
PeI test region
URA3
Centromere
Telomere
Southern probe for Amp^R
PCR amplicon for Amp^R
PCR amplicon for URA3
B
20 kb
10 kb
7 kb
5 kb
4 kb
3 kb
2 kb
C
Amp^R
URA3

Figure S11. Yeast artificial chromosome breakage assay detects chromosome breakage.

(A) To-scale diagram of a full length yeast artificial chromosome. Hash marks, gaps. *LEU2* and *URA3*, marker genes. Seed, telomere seed site. Amp^R, ampicillin resistance gene. EcoRI cuts between *LEU2* and Amp^R. Distance from EcoRI cut to end of telomere is ~12 kb. **(B)** Southern blot, hybridized with an Amp^R probe, of EcoRI-digested genomic DNA from a full length yeast artificial chromosome strain (leftmost lane) and 8 independent FOA^R leu+ colonies post-breakage (other lanes). Diagrams on right depict models of blotted DNA fragments at their respective fragment sizes (left-most gray box is Amp^R). **(C)** PCR of Amp^R and *URA3* amplicons using the same DNA as in (B).

Table S1.

Mutations underlying adaptive changes in recent human evolution. An extensive literature review identifies multiple loci thought to have contributed to adaptive evolution during the time frame of recent migration and expansion of human populations out of Africa (*56*). Although most currently known examples are limited to coding region changes (*56*), we note that half of the changes (47/94) likely arose by elevated mutation rate mechanisms, including 38 C to T mutations at CpG sites, plus 9 homopolymer slippage or large deletions/duplications. The high frequency of CpG transitions among the human mutations (47.6% (34/71) of single nucleotide changes occurring in coding regions) represents a ~17-fold enrichment over the ~2.8% frequency of CpG sites in human coding sequence (*60*)). DNA fragility at the *Pel* region and our population genetics modeling suggest that many currently unidentified regulatory mutations underlying other adaptive traits in humans (*61*) and sticklebacks (*31*) may also arise by high mutation rate mechanisms, a possibility that can be further tested as additional causative regulatory changes are found.

Table S1

Trait	Gene	Amino acid change	Nucleotide change	Mutation type	Variant ID	References (PubMed ID)
Alcohol response	ADH1B	R48H	C -> T	CpG	rs1229984	17847010, 18382665 20089146, 23071458
Alcohol response	ADH1B	R370C	C -> T	CpG	rs2066702	17847010, 18382665 20089146, 23071458
Blood type	ABO HBGG	Frameshift*	G258 deletion	Small indel	rs8176719	2121736, 18629539
Blood type	ABO HBGG	Frameshift*	G258 deletion	Small indel	rs8176719	2121736, 18629539
Blood type	ABO HBGG	Frameshift	G insertion in GGGGGGG795_801	Homopolymer	rs782782485	15647021, 18832934
Blood type	ABO HBGG	G268R	G -> A	SNP	rs41302905	8484251, 18832934
Blood type	ABO HBGG	Y309stop	C -> A	SNP	rs782586438	15647021, 18832934
Blood type	FUT2	A93T	G -> A	CpG	rs112722916	18422843
Blood type	FUT2	P112L	C -> T	CpG	rs200157007	10980544
Blood type	FUT2	I140F	A -> T	SNP	rs1047781	8526839, 19487333
Blood type	FUT2	W154stop	G -> A	SNP	rs601338	7876235, 19487333
Blood type	FUT2	R201H	G -> A	CpG	rs572832908	18422843
Blood type	FUT2	R202stop	C -> T	CpG	rs1800028	11404338, 8670215
Blood type	FUT2	R221stop	C -> T	CpG	rs1800029	11404338
Blood type	FUT2	R231stop*	C -> T	CpG	rs144566043	10319583
Blood type	FUT2	R231stop*	C -> T	CpG	rs144566043	17089126
Blood type	FUT2	R233C	C -> T	CpG	rs768236330	14569463
Blood type	FUT2	V240M	G -> A	CpG	rs375360260	10980544
Blood type	FUT2	Val241 inframe deletion	GTG718_720 deletion	Small indel	rs761766024	10980544
Blood type	FUT2	G247S	G -> A	CpG	rs602662	15250822
Blood type	FUT2	D254N	G -> A	SNP	rs768143018	14569463
Blood type	FUT2	Frameshift	C811 deletion	Small indel	rs1799761	9760207
Blood type	FUT2	T284N	C -> A	SNP	rs371279676	22188519
Blood type	FUT2	F291S	T -> C	SNP	CM042989	15250822
Blood type	FUT2	W294stop	G -> A	SNP	rs1800030	8670215
Blood type	FUT2	A296T	G -> A	CpG	rs79097987	22188519
Blood type	FUT2	G290R	G -> A	CpG	rs144269088	14569463
Blood type	FUT2	P317L	C -> T	CpG	rs200626231	18422843
Blood type	FUT2	Exon 2 deletion	Alu-mediated 10 kb deletion	Deletion		10982186, 19487333
Blood type	FUT2	Exon 2 deletion	Alu-mediated 9.3 kb deletion	Deletion		10980544, 19487333
Blood type	FUT2	Exon 2 partial deletion	Alu-mediated 4 kb deletion	Deletion		20880207
Blood type	FUT2	Gene fusion	Alu-mediated 22 kb deletion	Deletion		8755920
Ear wax type	ABCC11	G180R	G -> A	CpG	rs17822931	16444273, 19710689
Glycemia and adipolysis regulation	GIP	S103G	A -> G	SNP	rs2291725	20978139
Hair thickness	EDAR	V370A	T -> C	SNP	rs3827760	18065779
Immune system	IL4	Noncoding	C -> T	SNP	rs2243250	14654003
Lactose tolerance	LCT	Noncoding	C-13907G	SNP	rs41525747	17159977
Lactose tolerance	LCT	Noncoding	C-13910T	SNP	rs4988235	11788828, 17159977
Lactose tolerance	LCT	Noncoding	T-13915G	SNP	rs41380347	17159977
Lactose tolerance	LCT	Noncoding	C-14010G	CpG	rs145946881	17159977
Masticatory muscles	MYH16	Frameshift	Exon 18 AC deletion	Small indel		15042088
Metabolism (drugs)	CYP2C9	R144C	C -> T	CpG	rs1799853	8873220, 19300499 20833655, 27803446
Metabolism (drugs)	CYP2C9	I359L	A -> C	SNP	rs1057910	8873220, 19300499 20833655, 27803446
Metabolism (environmental toxins)	CYP2C19	Noncoding	C-806T	SNP	rs12248560	16413245, 24690327
Metabolism (environmental toxins)	CYP2C19	W212stop	G -> A	SNP	rs4986893	24690327
Metabolism (environmental toxins)	CYP2C19	P227P (silent)	G -> A	CpG	rs4244285	24690327
Metabolism (warfarin)	CYP4F2	V433M	G -> A	SNP	rs2108622	19300499, 20833655 14765194, 15888487
Metabolism (warfarin)	VKORC1	Noncoding	G-1639A	CpG	rs9923231	18252229, 19300499 20555338, 20833655 14765194, 15888487
Metabolism (warfarin)	VKORC1	V29L	G -> T	CpG	rs104894539	18252229, 19300499 20555338, 20833655 14765194, 15888487
Metabolism (warfarin)	VKORC1	D36Y	G -> T	SNP	rs61742245	18252229, 19300499 20555338, 20833655 14765194, 15888487
Metabolism (warfarin)	VKORC1	V45A	T -> C	SNP	rs104894540	18252229, 19300499 20555338, 20833655 14765194, 15888487
Metabolism (warfarin)	VKORC1	R58G	A -> G	SNP	rs104894541	18252229, 19300499 20555338, 20833655 14765194, 15888487
Metabolism (warfarin)	VKORC1	V66M	G -> A	SNP	rs72547529	18252229, 19300499 20555338, 20833655 14765194, 15888487
Metabolism (warfarin)	VKORC1	R98W*	C -> T	CpG	rs72547528	18252229, 19300499 20555338, 20833655 14765194, 15888487
Metabolism (warfarin)	VKORC1	R98W*	C -> T	CpG	rs72547528	18252229, 19300499 20555338, 20833655

Metabolism (warfarin)	VKORC1	L128R	T -> G	SNP	rs104894542	14765194, 15888487 18252229, 19300499
Neuronal maturation	SRGAP2B	Gene duplication	258 kb interspersed duplication	Duplication		20555338, 20833655 22559943, 22559944
Neuronal maturation	SRGAP2C	Gene duplication	>515 kb interspersed duplication	Duplication		22559943, 22559944
Olfaction	OR7D4	P79L	C -> T	SNP	rs61732668	17873857, 19955411
Olfaction	OR7D4	N84S	A -> G	SNP	rs5020280	17873857, 19955411
Olfaction	OR7D4	R88W	C -> T	CpG	rs61729907	17873857, 19955411
Olfaction	OR7D4	T133M	C -> T	CpG	rs5020278	17873857, 19955411
Pain response	COMT	H62H (silent)	C -> T	CpG	rs4633	17185601
Pain response	COMT	L136L (silent)	G -> C	SNP	rs4818	17185601
Pain response	COMT	V158M	G -> A	CpG	rs4680	17185601
Pathogen response (HIV)	CCL3L1	Gene duplication	>14 kb tandem duplication	Duplication		15637236
Pathogen response (malaria)	DARC	Noncoding	T-46C	SNP	rs1800846	7663520
Pathogen response (malaria)	HBB	E6V	A -> T	SNP	rs334	19465909
Pathogen response (malaria)	G6PD	H32R	A -> G	SNP	rs137852340	17978087, 16607506
Pathogen response (malaria)	G6PD	A44G	C -> G	SNP	rs78478128	17978087, 8533762
Pathogen response (malaria)	G6PD	V68M	G -> A	CpG	rs1050828	17978087, 11423617 16020776
Pathogen response (malaria)	G6PD	Y70H	T -> C	SNP	rs137852349	17233850
Pathogen response (malaria)	G6PD	N126D	A -> G	SNP	rs1050829	17978087, 11423617
Pathogen response (malaria)	G6PD	L128P	T -> C	SNP	rs78365220	17978087
Pathogen response (malaria)	G6PD	G131V	G -> T	SNP	rs137852341	16607506
Pathogen response (malaria)	G6PD	G163S	G -> A	SNP	rs137852314	17978087, 20007901
Pathogen response (malaria)	G6PD	D181V	A -> T	SNP	rs5030872	17978087, 12367584 1999409, 16461316
Pathogen response (malaria)	G6PD	S188F	C -> T	SNP	rs5030868	17978087, 11423617
Pathogen response (malaria)	G6PD	R198C	C -> T	CpG	COSM3559803	1551674, 16607506
Pathogen response (malaria)	G6PD	R198H	G -> A	CpG	rs782583168	18043863
Pathogen response (malaria)	G6PD	D282H	G -> C	SNP	rs137852318	17978087, 12367584
Pathogen response (malaria)	G6PD	V291M	G -> A	SNP	rs137852327	17978087, 16136268
Pathogen response (malaria)	G6PD	E317K	G -> A	CpG	rs137852339	17978087, 5673160
Pathogen response (malaria)	G6PD	L323P	T -> C	SNP	rs76723693	17978087, 16461316
Pathogen response (malaria)	G6PD	A335T	G -> A	CpG	rs5030869	17978087, 12028056
Pathogen response (malaria)	G6PD	R454H	G -> A	CpG	rs137852324	2393028, 16088936
Pathogen response (malaria)	G6PD	R454C	C -> T	CpG	rs398123546	16607506, 16088936
Pathogen response (malaria)	G6PD	R459L	G -> T	CpG	rs72554665	17978087, 16607506
Pathogen response (malaria)	G6PD	R459P	G -> C	SNP	rs72554665	12028056, 16143877
Pathogen response (malaria)	G6PD	R463H	G -> A	CpG	rs72554664	17978087, 16607506
Pigmentation (eyes and skin)	OCA2	Noncoding	A -> G	SNP	rs12913832	18172690, 18483556 22234890
Pigmentation (skin)	SLC24A5	A111T	G -> A	CpG	rs1426654	16357253, 16524431 18166528, 22198722
Pigmentation (hair)	TYRP1	R93C	C -> T	CpG	rs387907171	22556244
Starch digestion	AMY1	Gene duplication	>18 kb tandem duplication	Duplication		17828263

* Likely arose multiple times independently

Table S2.
Overlap of ecology-associated CNVs with different types of dinucleotide repeats.

Table S2

EcoCNV type	Total	Near TG repeats		Near TC repeats		Near TA repeats	
		Number	Percent	Number	Percent	Number	Percent
All	6664	693	10.4%	90	1.4%	221	3.3%
Deletions	659	98	14.9% *	16	2.4%	19	2.9%
Insertions	5224	521	10.0%	67	1.3%	165	3.2%
Other	781	74	9.5%	7	0.9%	37	4.7%

"Near" defined as overlapping or within 1 kb
* *P* vs. all CNVs < 9.7e-5

Table S3.

Yeast strains used and generated in this study.

Table S3

Name	Background	Genotype	YAC genotype	Reference
CFY1700	S288C BY4741	*MATa leu2Δ0 ura3Δ0 met15Δ0 bar1Δ::KAN*	*LEU2*	Gift from C. Freudenreich
KXY1	CFY1700	*MATa leu2Δ0 ura3Δ0 met15Δ0 bar1Δ::KAN*	*LEU2 VS20N-Empty URA3*	This study
KXY61	CFY1700	*MATa leu2Δ0 ura3Δ0 met15Δ0 bar1Δ::KAN*	*LEU2 VS20N-Empty URA3*	This study
KXY111	CFY1700	*MATa leu2Δ0 ura3Δ0 met15Δ0 bar1Δ::KAN*	*LEU2 VS20N-Pel-RABS URA3*	This study
KXY118	CFY1700	*MATa leu2Δ0 ura3Δ0 met15Δ0 bar1Δ::KAN*	*LEU2 VS20N-Pel-RABS URA3*	This study
KXY6	CFY1700	*MATa leu2Δ0 ura3Δ0 met15Δ0 bar1Δ::KAN*	*LEU2 VS20N-Pel-LITC URA3*	This study
KXY67	CFY1700	*MATa leu2Δ0 ura3Δ0 met15Δ0 bar1Δ::KAN*	*LEU2 VS20N-Pel-LITC URA3*	This study
KXY87	CFY1700	*MATa leu2Δ0 ura3Δ0 met15Δ0 bar1Δ::KAN*	*LEU2 VS20N-Pel-BDGB URA3*	This study
KXY88	CFY1700	*MATa leu2Δ0 ura3Δ0 met15Δ0 bar1Δ::KAN*	*LEU2 VS20N-Pel-BDGB URA3*	This study
KXY14	CFY1700	*MATa leu2Δ0 ura3Δ0 met15Δ0 bar1Δ::KAN*	*LEU2 VS20N-Pel-TOAD URA3*	This study
KXY74	CFY1700	*MATa leu2Δ0 ura3Δ0 met15Δ0 bar1Δ::KAN*	*LEU2 VS20N-Pel-TOAD URA3*	This study
KXY43	CFY1700	*MATa leu2Δ0 ura3Δ0 met15Δ0 bar1Δ::KAN*	*LEU2 VS20N-Pel-PAXB URA3*	This study
KXY93	CFY1700	*MATa leu2Δ0 ura3Δ0 met15Δ0 bar1Δ::KAN*	*LEU2 VS20N-Pel-PAXB URA3*	This study
KXY112	CFY1700	*MATa leu2Δ0 ura3Δ0 met15Δ0 bar1Δ::KAN*	*LEU2 VS20N-Pel-RABS(RC) URA3*	This study
KXY113	CFY1700	*MATa leu2Δ0 ura3Δ0 met15Δ0 bar1Δ::KAN*	*LEU2 VS20N-Pel-RABS(RC) URA3*	This study
KXY110	CFY1700	*MATa leu2Δ0 ura3Δ0 met15Δ0 bar1Δ::KAN*	*LEU2 VS20N-Pel-LITC(RC) URA3*	This study
KXY123	CFY1700	*MATa leu2Δ0 ura3Δ0 met15Δ0 bar1Δ::KAN*	*LEU2 VS20N-Pel-LITC(RC) URA3*	This study
KXY96	CFY1700	*MATa leu2Δ0 ura3Δ0 met15Δ0 bar1Δ::KAN*	*LEU2 VS20N-Pel-BDGB(RC) URA3*	This study
KXY97	CFY1700	*MATa leu2Δ0 ura3Δ0 met15Δ0 bar1Δ::KAN*	*LEU2 VS20N-Pel-BDGB(RC) URA3*	This study
KXY179	CFY1700	*MATa leu2Δ0 ura3Δ0 met15Δ0 bar1Δ::KAN*	*LEU2 VS20NR-Empty URA3*	This study
KXY180	CFY1700	*MATa leu2Δ0 ura3Δ0 met15Δ0 bar1Δ::KAN*	*LEU2 VS20NR-Empty URA3*	This study
KXY182	CFY1700	*MATa leu2Δ0 ura3Δ0 met15Δ0 bar1Δ::KAN*	*LEU2 VS20NR-Pel-RABS URA3*	This study
KXY183	CFY1700	*MATa leu2Δ0 ura3Δ0 met15Δ0 bar1Δ::KAN*	*LEU2 VS20NR-Pel-RABS URA3*	This study
KXY187	CFY1700	*MATa leu2Δ0 ura3Δ0 met15Δ0 bar1Δ::KAN*	*LEU2 VS20NR-Pel-RABS(RC) URA3*	This study
KXY189	CFY1700	*MATa leu2Δ0 ura3Δ0 met15Δ0 bar1Δ::KAN*	*LEU2 VS20NR-Pel-RABS(RC) URA3*	This study
KXY191	CFY1700	*MATa leu2Δ0 ura3Δ0 met15Δ0 bar1Δ::KAN*	*LEU2 VS20N-Empty(+ori) URA3*	This study
KXY192	CFY1700	*MATa leu2Δ0 ura3Δ0 met15Δ0 bar1Δ::KAN*	*LEU2 VS20N-Empty(+ori) URA3*	This study
KXY121	CFY1700	*MATa leu2Δ0 ura3Δ0 met15Δ0 bar1Δ::KAN*	*LEU2 VS20N-Pel-RABS(+ori) URA3*	This study
KXY122	CFY1700	*MATa leu2Δ0 ura3Δ0 met15Δ0 bar1Δ::KAN*	*LEU2 VS20N-Pel-RABS(+ori) URA3*	This study
KXY128	CFY1700	*MATa leu2Δ0 ura3Δ0 met15Δ0 bar1Δ::KAN*	*LEU2 VS20N-Pel-RABS(RC+ori) URA3*	This study
KXY129	CFY1700	*MATa leu2Δ0 ura3Δ0 met15Δ0 bar1Δ::KAN*	*LEU2 VS20N-Pel-RABS(RC+ori) URA3*	This study
KXY108	CFY1700	*MATa leu2Δ0 ura3Δ0 met15Δ0 bar1Δ::KAN*	*LEU2 VS20N-$(TG)_{14}$ URA3*	This study
KXY125	CFY1700	*MATa leu2Δ0 ura3Δ0 met15Δ0 bar1Δ::KAN*	*LEU2 VS20N-$(TG)_{14}$ URA3*	This study
KXY126	CFY1700	*MATa leu2Δ0 ura3Δ0 met15Δ0 bar1Δ::KAN*	*LEU2 VS20N-$(TG)_{43}$ URA3*	This study
KXY127	CFY1700	*MATa leu2Δ0 ura3Δ0 met15Δ0 bar1Δ::KAN*	*LEU2 VS20N-$(TG)_{43}$ URA3*	This study
KXY109	CFY1700	*MATa leu2Δ0 ura3Δ0 met15Δ0 bar1Δ::KAN*	*LEU2 VS20N-$(TG)_{79}$ URA3*	This study
KXY120	CFY1700	*MATa leu2Δ0 ura3Δ0 met15Δ0 bar1Δ::KAN*	*LEU2 VS20N-$(TG)_{79}$ URA3*	This study
KXY106	CFY1700	*MATa leu2Δ0 ura3Δ0 met15Δ0 bar1Δ::KAN*	*LEU2 VS20N-$(CA)_{16}$ URA3*	This study
KXY107	CFY1700	*MATa leu2Δ0 ura3Δ0 met15Δ0 bar1Δ::KAN*	*LEU2 VS20N-$(CA)_{16}$ URA3*	This study
KXY104	CFY1700	*MATa leu2Δ0 ura3Δ0 met15Δ0 bar1Δ::KAN*	*LEU2 VS20N-$(CA)_{50}$ URA3*	This study
KXY105	CFY1700	*MATa leu2Δ0 ura3Δ0 met15Δ0 bar1Δ::KAN*	*LEU2 VS20N-$(CA)_{50}$ URA3*	This study

Table S4.
Detailed yeast artificial chromosome assay breakage rates.

Table S4

Ecotype	YAC context	Test sequence	Breaks per division	Breaks per division (construct average)	Fold above empty* (construct average)	Figure reference
		Empty vector	6.42E-06	3.37E-06	1.00	2B, 2C, 3C
		"	2.99E-06	"	"	"
		"	1.79E-08	"	"	"
		"	8.47E-07	"	"	"
		"	4.05E-06	"	"	"
		"	5.90E-06	"	"	"
Marine		BDGB	6.32E-05	9.54E-05	28.31	2B
"		"	1.01E-04	"	"	"
"		"	3.61E-05	"	"	"
"		"	5.15E-05	"	"	"
"		"	1.69E-04	"	"	"
"		"	1.52E-04	"	"	"
Marine		LITC	2.46E-04	1.94E-04	57.70	2B
"		"	2.30E-04	"	"	"
"		"	1.31E-04	"	"	"
"		"	1.37E-04	"	"	"
"		"	1.80E-04	"	"	"
"		"	2.43E-04	"	"	"
Marine		RABS	1.15E-04	1.15E-04	34.12	2B, 2C
"		"	1.09E-04	"	"	"
"		"	7.82E-05	"	"	"
"		"	4.64E-05	"	"	"
"		"	8.45E-05	"	"	"
"		"	2.57E-04	"	"	"
Freshwater		TOAD	7.96E-06	8.05E-06	2.39	2B
"		"	9.92E-06	"	"	"
"		"	9.35E-06	"	"	"
"		"	6.41E-06	"	"	"
"		"	7.92E-06	"	"	"
"		"	6.75E-06	"	"	"
Freshwater		PAXB	7.61E-07	3.01E-06	0.89	2B
"		"	1.01E-06	"	"	"
"		"	3.38E-06	"	"	"
"		"	1.53E-06	"	"	"
"		"	1.36E-06	"	"	"
"		"	1.00E-05	"	"	"
Marine		BDGB RC	3.69E-06	1.13E-05	3.34	2B
"		"	5.74E-06	"	"	"
"		"	1.01E-05	"	"	"

"		"	2.17E-05	"	"	"
"		"	2.33E-05	"	"	"
"		"	3.09E-06	"	"	"
Marine		LITC RC	8.91E-06	9.77E-06	2.90	2B
"		"	3.44E-06	"	"	"
"		"	4.97E-06	"	"	"
"		"	2.25E-06	"	"	"
"		"	2.15E-05	"	"	"
"		"	1.75E-05	"	"	"
Marine		RABS RC	4.07E-06	7.06E-06	2.09	2B, 2C
"		"	9.28E-06	"	"	"
"		"	7.47E-06	"	"	"
"		"	1.12E-06	"	"	"
"		"	2.91E-06	"	"	"
"		"	1.75E-05	"	"	"
	Reversed transcription	Empty vector	3.65E-06	8.29E-06	1.00	2C
	"	"	6.83E-06	"	"	"
	"	"	2.16E-06	"	"	"
	"	"	9.17E-06	"	"	"
	"	"	1.01E-05	"	"	"
	"	"	1.78E-05	"	"	"
Marine	Reversed transcription	RABS	1.48E-04	1.99E-04	23.96	2C
"	"	"	2.90E-04	"	"	"
"	"	"	3.59E-04	"	"	"
"	"	"	1.55E-04	"	"	"
"	"	"	9.93E-05	"	"	"
"	"	"	1.41E-04	"	"	"
Marine	Reversed transcription	RABS RC	8.50E-06	1.37E-05	1.65	2C
"	"	"	9.17E-06	"	"	"
"	"	"	1.09E-05	"	"	"
"	"	"	1.41E-05	"	"	"
"	"	"	1.14E-05	"	"	"
"	"	"	2.82E-05	"	"	"
	Reversed replication	Empty vector	7.98E-07	1.43E-06	1.00	2C
	"	"	1.18E-06	"	"	"
	"	"	2.14E-06	"	"	"
	"	"	9.81E-07	"	"	"
	"	"	6.01E-07	"	"	"
	"	"	2.88E-06	"	"	"
Marine	Reversed replication	RABS	7.04E-07	1.06E-06	0.74	2C

"	"	"	9.35E-07	"	"	"
"	"	"	3.50E-07	"	"	"
"	"	"	2.10E-06	"	"	"
"	"	"	1.61E-06	"	"	"
"	"	"	6.52E-07	"	"	"
Marine	Reversed replication	RABS RC	5.26E-05	2.37E-05	16.55	2C
"	"	"	1.99E-05	"	"	"
"	"	"	1.41E-05	"	"	"
"	"	"	1.70E-05	"	"	"
"	"	"	1.50E-05	"	"	"
"	"	"	2.33E-05	"	"	"
		(TG)14	9.25E-06	1.57E-05	4.67	3C
		"	1.90E-05	"	"	"
		"	5.34E-06	"	"	"
		"	1.16E-05	"	"	"
		"	1.86E-05	"	"	"
		"	3.06E-05	"	"	"
		(TG)43	1.16E-04	1.04E-04	30.90	3C
		"	1.66E-04	"	"	"
		"	1.19E-04	"	"	"
		"	9.27E-05	"	"	"
		"	6.01E-05	"	"	"
		"	7.05E-05	"	"	"
		(TG)79	1.34E-04	1.35E-04	40.14	3C
		"	1.69E-04	"	"	"
		"	9.58E-05	"	"	"
		"	1.34E-04	"	"	"
		"	1.69E-04	"	"	"
		"	1.10E-04	"	"	"
		(CA)16	7.27E-06	5.55E-06	1.65	3C
		"	4.72E-06	"	"	"
		"	1.13E-06	"	"	"
		"	4.41E-06	"	"	"
		"	6.59E-06	"	"	"
		"	9.16E-06	"	"	"
		(CA)50	2.63E-06	4.89E-06	1.45	3C
		"	1.45E-06	"	"	"
		"	6.06E-06	"	"	"
		"	1.29E-06	"	"	"
		"	9.81E-06	"	"	"
		"	8.09E-06	"	"	"
Freshwater intact pelvis		NNCY	3.08E-04	2.92E-04	86.49	S7B
"		"	2.89E-04	"	"	"

"	"	1.81E-04	"	"	"
"	"	4.47E-04	"	"	"
"	"	1.71E-04	"	"	"
"	"	3.53E-04	"	"	"
Freshwater intact pelvis	MAYR	3.34E-04	2.19E-04	65.12	S7B
"	"	8.54E-05	"	"	"
"	"	2.78E-04	"	"	"
"	"	2.54E-04	"	"	"
"	"	8.54E-05	"	"	"
"	"	2.80E-04	"	"	"
Freshwater intact pelvis	MATA	2.37E-04	1.66E-04	49.19	S7B
"	"	6.93E-05	"	"	"
"	"	2.15E-04	"	"	"
"	"	2.09E-04	"	"	"
"	"	6.36E-05	"	"	"
"	"	2.01E-04	"	"	"

* Calculated as (construct average)/(empty vector average)
except for reversed transcription/replication constructs, where the paired empty vector construct average was used

RC = Reverse complement

Table S5.
Detailed yeast artificial chromosome assay p-value comparisons.

Table S5

Two sample tests vs. Empty vector	Wilcoxon	*t* test
Empty vs Marine*	0.000015	0.0002
Empty vs Freshwater*	0.179703	0.2198
Empty vs Marine RC*	0.077220	0.0715
Empty vs BDGB	0.002165	0.0022
Empty vs LITC	0.002165	< 0.0001
Empty vs RABS	0.002165	0.0041
Empty vs TOAD**	0.004329	0.0030
Empty vs PAXB	0.818182	0.8436
Empty vs BDGB RC	0.132035	0.0671
Empty vs LITC RC	0.179654	0.0915
Empty vs RABS RC	0.240260	0.1931
Empty vs (TG)14	0.008658	0.0091
Empty vs (TG)43	0.002165	< 0.0001
Empty vs (TG)79	0.004922	< 0.0001
Empty vs (CA)16	0.132035	0.1919
Empty vs (CA)50	0.484848	0.4244
Reverse transcription Empty vs Reverse transcription RABS	0.002165	0.0010
Reverse transcription Empty vs Reverse transcription RABS RC	0.148829	0.1808
Reverse replication Empty vs Reverse replication RABS	0.393939	0.4312
Reverse replication Empty vs Reverse replication RABS RC	0.002165	0.0039
Empty vs Freshwater intact pelvis*	0.000360	< 0.0001
Empty vs NNCY	0.002165	< 0.0001
Empty vs MAYR	0.004998	0.0006
Empty vs MATA	0.002165	0.0005

Multiple sample tests with post-hoc multiple comparison***	Kruskal-Wallis****	ANOVA ****
Fig. 2B (Empty, Marine, Freshwater, Marine RC)	**4.084781e-08**	**1.3504e-12**
Empty vs Marine*	0.000016	0.0010053
Empty vs Freshwater*	0.448108	0.8999947
Empty vs Marine RC*	0.213415	0.8999947
Marine* vs. Freshwater*	0.000010	0.0010053
Marine* vs. Marine RC*	0.000016	0.0010053
Freshwater* vs. Marine RC*	0.472054	0.8999947
Conclusion: Marine is significantly different from rest of group		
Fig. 2C (Empty, RABS, RABS RC)	**0.002338**	**0.0005**
Empty vs RABS	0.002402	0.0011008
Empty vs RABS RC	0.386940	0.8999947
RABS vs RABS RC	0.025736	0.0014785
Conclusion: RABS is significantly different from rest of group		
Fig. 2C (Reverse transcription: Empty, RABS, RABS RC)	**0.001992**	**5.4254e-05**
Reverse transcription Empty vs Reverse transcription RABS	0.001775	0.0010053
Reverse transcription Empty vs Reverse transcription RABS RC	0.303981	0.8999947
Reverse transcription RABS vs. Reverse transcription RABS RC	0.032122	0.0010053
Conclusion: Reverse transcription RABS is significantly different from rest of group		
Fig. 2C (Reverse replication: Empty, RABS, RABS RC)	**0.002754**	**0.0004**

Reverse replication Empty vs Reverse replication RABS	0.516412	0.8999947
Reverse replication Empty vs Reverse replication RABS RC	0.014166	0.0010270
Reverse replication RABS vs Reverse replication RABS RC	0.003531	0.0010053
Conclusion: Reverse replication RABS RC is significantly different from rest of group		

All calculations were made using the median values reported in table S5
Grey indicates not significant ($p > 0.05$)

* Marine indicates median values from BDGB, LITC, and RABS combined
Freshwater indicates median values from TOAD and PAXB combined
Marine RC indicates median values from BDGB RC, LITC RC, and RABS RC combined
Freshwater intact pelvis indicates median values from NNCY, MAYR, and MATA combined

** The TOAD deletion allele retains a (TG)15 stretch (file S1)

*** Comparison groupings are indicated in bold; bolded p-values indicate overall p-value without post-hoc analysis; unbolded p-values indicate p-values from post-hoc pairwise multiple comparison

**** Kruskal-Wallis with post-hoc Dunn p-values, further adjusted by the Benjamini-Hochberg False Discovery Rate method
ANOVA with post-hoc Tukey Honest Significant Difference p-values

Table S6.
Stickleback populations used in this study.

Table S6

Acronym	Population	Ecology	Pelvic phenotype	Location	N Latitude	W Longitude
BDGB	Bodega Bay	Marine	Full	USA, California	38.325	123.041
BEPA	Bear Paw Lake	Freshwater	Reduced	USA, Alaska	61.614	149.756
BOOT	Boot Lake	Freshwater	Reduced	USA, Alaska	61.717	150.117
BOUL	Boulton Lake	Freshwater	Reduced	Canada, British Columbia	53.783	132.098
CMCB	Community Club Pond	Freshwater	Reduced	USA, Alaska	60.702	151.383
HUMP	Hump Lake	Freshwater	Reduced	USA, Alaska	60.769	151.167
JADE	Jade Lake	Freshwater	Full	USA, Alaska	61.524	149.869
KFSY	Kalifonsky Lake	Freshwater	Reduced	USA, Alaska	60.331	151.264
LITC	Little Campbell River	Marine	Full	Canada, British Columbia	49.018	122.779
LSHP	L-Shaped Lake	Freshwater	Reduced	USA, Alaska	61.706	149.972
MATA	Matadero	Freshwater	Full	USA, California	37.386	122.165
MAYR	Mayer	Freshwater	Full	Canada, British Columbia	53.644	132.057
NNCY	Nancy	Freshwater	Full	USA, Alaska	61.685	150.000
ORPH	Orphea Lake	Freshwater	Reduced	USA, Alaska	60.386	151.200
PAXB	Paxton Lake (Benthic)	Freshwater	Reduced	Canada, British Columbia	49.712	124.525
RABS	Rabbit Slough	Marine	Full	USA, Alaska	61.537	149.166
TOAD	Toad Lake	Freshwater	Reduced	USA, Alaska	61.619	149.696