doi:10.1038/nature10944

The genomic basis of adaptive evolution in threespine sticklebacks

Felicity C. Jones[1]*, Manfred G. Grabherr[2,3]*, Yingguang Frank Chan[1]†*, Pamela Russell[2]*, Evan Mauceli[2]†, Jeremy Johnson[2], Ross Swofford[2], Mono Pirun[2]†, Michael C. Zody[2], Simon White[4], Ewan Birney[5], Stephen Searle[4], Jeremy Schmutz[6], Jane Grimwood[6], Mark C. Dickson[6], Richard M. Myers[6], Craig T. Miller[1]†, Brian R. Summers[1], Anne K. Knecht[1], Shannon D. Brady[1], Haili Zhang[1], Alex A. Pollen[1], Timothy Howes[1], Chris Amemiya[7], Broad Institute Genome Sequencing Platform & Whole Genome Assembly Team‡, Eric S. Lander[2], Federica Di Palma[2], Kerstin Lindblad-Toh[2,3] & David M. Kingsley[1,8]

Marine stickleback fish have colonized and adapted to thousands of streams and lakes formed since the last ice age, providing an exceptional opportunity to characterize genomic mechanisms underlying repeated ecological adaptation in nature. Here we develop a high-quality reference genome assembly for threespine sticklebacks. By sequencing the genomes of twenty additional individuals from a global set of marine and freshwater populations, we identify a genome-wide set of loci that are consistently associated with marine–freshwater divergence. Our results indicate that reuse of globally shared standing genetic variation, including chromosomal inversions, has an important role in repeated evolution of distinct marine and freshwater sticklebacks, and in the maintenance of divergent ecotypes during early stages of reproductive isolation. Both coding and regulatory changes occur in the set of loci underlying marine–freshwater evolution, but regulatory changes appear to predominate in this well known example of repeated adaptive evolution in nature.

The genetic and molecular basis of adaptive evolution is still largely unknown. Some researchers have championed a pre-eminent role for regulatory changes during evolution of adaptive phenotypes, because such changes may avoid pleiotropic consequences of protein-coding alterations[1–3]. Others have catalogued known phenotypic differences caused by protein-coding changes and have questioned whether sufficient case histories exist to estimate the relative frequency of regulatory and coding changes during adaptive evolution[4]. Despite progress on individual traits[5], it has been difficult to accumulate enough examples in any particular group to obtain an overall picture of molecular mechanisms underlying evolutionary change, particularly for clearly adaptive phenotypes in wild organisms.

Threespine sticklebacks offer a powerful system for studying the molecular basis of adaptive evolution in vertebrates. After the retreat of Pleistocene glaciers, marine sticklebacks colonized and adapted to many newly formed freshwater habitats, evolving repeated changes in body shape, skeletal armour, trophic specializations, pigmentation, salt handling, life history and mating preferences[6,7]. Recurrent evolution of similar phenotypes in similar environments indicates that these traits evolve by natural selection[8]. Distinctive marine and freshwater forms can still hybridize, making it possible to map the genetic basis of individual traits, and identify particular genes underlying armour, pelvic and pigmentation evolution[9–12]. At two of these key loci, distinctive haplotypes were found to be reused when similar phenotypes evolve in different populations[11,12], a pattern that was later found at additional loci[13,14]. Ongoing gene flow between marine and freshwater forms occurs along coastal rivers[15,16], making it possible to spread adaptive alleles among populations, and homogenizing neutral genomic regions[17]. Here we use signatures of allele sharing to identify a genome-wide set of adaptive loci consistently associated with recurrent marine–freshwater evolution.

Generation of reference genome assembly

To facilitate studies of stickleback evolution, we first generated a reference genome assembly from a homogametic (female) freshwater stickleback (*Gasterosteus aculeatus*) from Bear Paw Lake, Alaska. The sequenced individual was partially inbred and retained heterozygosity at approximately 1 per 700 base pairs (bp). The assembly, gasAcu1.0, was generated with 9.0× coverage in Sanger sequence data (ABI3730), and has a length-weighted median (N50) contig size of 83.2 kilobases (kb), a length-weighted median (N50) scaffold size of 10.8 megabases (Mb) and a total gapped size of 463 Mb, close to previous estimates of 530 Mb (ref. 18). The 113 largest scaffolds (86.9%, 400.4 Mb) were anchored to stickleback linkage groups in an F_2 marine × freshwater intercross, whereas 60.7 Mb in 1,812 smaller scaffolds (N50 = 0.3 Mb) remain unanchored. Use of a single partially inbred individual, construction and assembly of a range of genomic library sizes, and the relatively low repeat and duplication content of the stickleback genome have produced a highly contiguous anchored genome assembly with contig and scaffold sizes much larger than other published teleosts[19–22] (Supplementary Table 1).

The stickleback sequence was annotated using the Ensembl pipeline, which predicted 20,787 protein-coding and 1,617 RNA genes (Supplementary Table 2). Of the protein-coding genes, 7,614 showed

[1]Department of Developmental Biology, Beckman Center B300, Stanford University School of Medicine, Stanford California 94305, USA. [2]Broad Institute of MIT and Harvard, 7 Cambridge Center, Cambridge Massachusetts 02142, USA. [3]Science for Life Laboratory Uppsala, Department of Medical Biochemistry and Microbiology, Uppsala University, Uppsala 751 23, Sweden. [4]Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SA, UK. [5]European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK. [6]HudsonAlpha Institute for Biotechnology, 601 Genome Way, Huntsville, Alabama 35806, USA. [7]Department of Molecular Genetics, Benaroya Research Institute at Virginia Mason, 1201 Ninth Avenue, Seattle Washington 98101, USA. [8]Howard Hughes Medical Institute, Stanford University, Stanford, California 94305, USA. †Present addresses: Max Planck Institute for Evolutionary Biology, August-Thienemann-Str. 2, Plön 24306, Germany (Y.F.C.); Children's Hospital Boston, Genetic Diagnostic Lab, 300 Longwood Avenue, Boston, Massachusetts 02115, USA (E.M.); Bioinformatics Core, Zuckerman Research Center, New York, New York 10065, USA (M.P.); Department of Molecular & Cell Biology, 142 LSA 3200, University of California, Berkeley, California 94720, USA (C.T.M.).

*These authors contributed equally to this work.

‡Lists of authors and affiliations appear at the end of the paper.

one-to-one orthology with mammals and an additional 7,192 showed one-to-one orthology among fishes. The other 5,981 genes showed complex orthology relationships, including some lineage-specific gene expansions that contribute to stickleback adaptations (for example, a duplicated mucin family encoding glue proteins used for male nest building[23]). A total of 13.4% of the stickleback genome appeared to be under evolutionary constraint when compared with other fishes using PhastCons[24]. The conserved portion was roughly equally divided between protein-coding and non-coding sequences, with ~71% of the latter shared with mammals and ~29% representing fish-specific conserved sequences (Supplementary Table 3).

Sequencing additional population pairs

To search for loci underlying repeated evolution in sticklebacks, we first identified populations showing characteristic marine and freshwater morphology (Fig. 1a, Supplementary Fig. 1 and Supplementary Table 4). Repeated adaptation to divergent marine and freshwater environments resulted in marked correlated changes in body shape, length, depth, fin position, spine length, eye size and armour plate number (Fig. 1b). Because quantitative trait loci (QTL) controlling these traits map to many different chromosomes[12,25–30], this morphological screen should identify populations differing in a genome-wide range of adaptive loci underlying marine–freshwater differences.

From the distinct morphological clusters of marine and freshwater fish, we selected multiple marine–freshwater pairs, from both Pacific and Atlantic populations, including individuals from opposite ends of rivers with marine–freshwater hybrid zones[16] (21 fish in total, including the reference genome individual). The sampling strategy should minimize geographic bias in the data set, while maximizing the chance for local exchange of neutral regions of the genome.

We generated 2.3× average coverage per individual using Illumina sequencing (Supplementary Table 5 and Supplementary Information). To identify single nucleotide polymorphisms (SNPs), we pooled data from all fish and identified positions where at least four reads support a variant allele. This criterion identified 5,897,368 candidate SNPs (Supplementary Table 6), with most being true positives based on experimental validation ($n = 48$ tested, 82.6% confirmed; Supplementary Information).

Genome-wide survey of parallel evolution

Previous studies have shown that repeated armour evolution in sticklebacks occurs through ancient variants at the *EDA* locus, which are reused in multiple freshwater populations[11] and are subject to strong selection[31]. To identify loci where alleles have similarly been used repeatedly during adaptive divergence of marine and freshwater fish, we used two methods to look for regions where sequences of most freshwater fish were similar to each other, but differed from sequences typically found in marine populations. Note that this pattern will not identify adaptive variants that are unique to individual freshwater populations, but instead focuses on variants with striking evidence of biological replication across populations.

First, we developed a self-organizing map-based iterative Hidden Markov Model (SOM/HMM) to identify the 20 most common patterns of genetic relationships ('trees') among the 21 individuals. Genomic regions were assigned to pattern types on the basis of likelihood, with boundaries defined using HMM transitions. This method iteratively

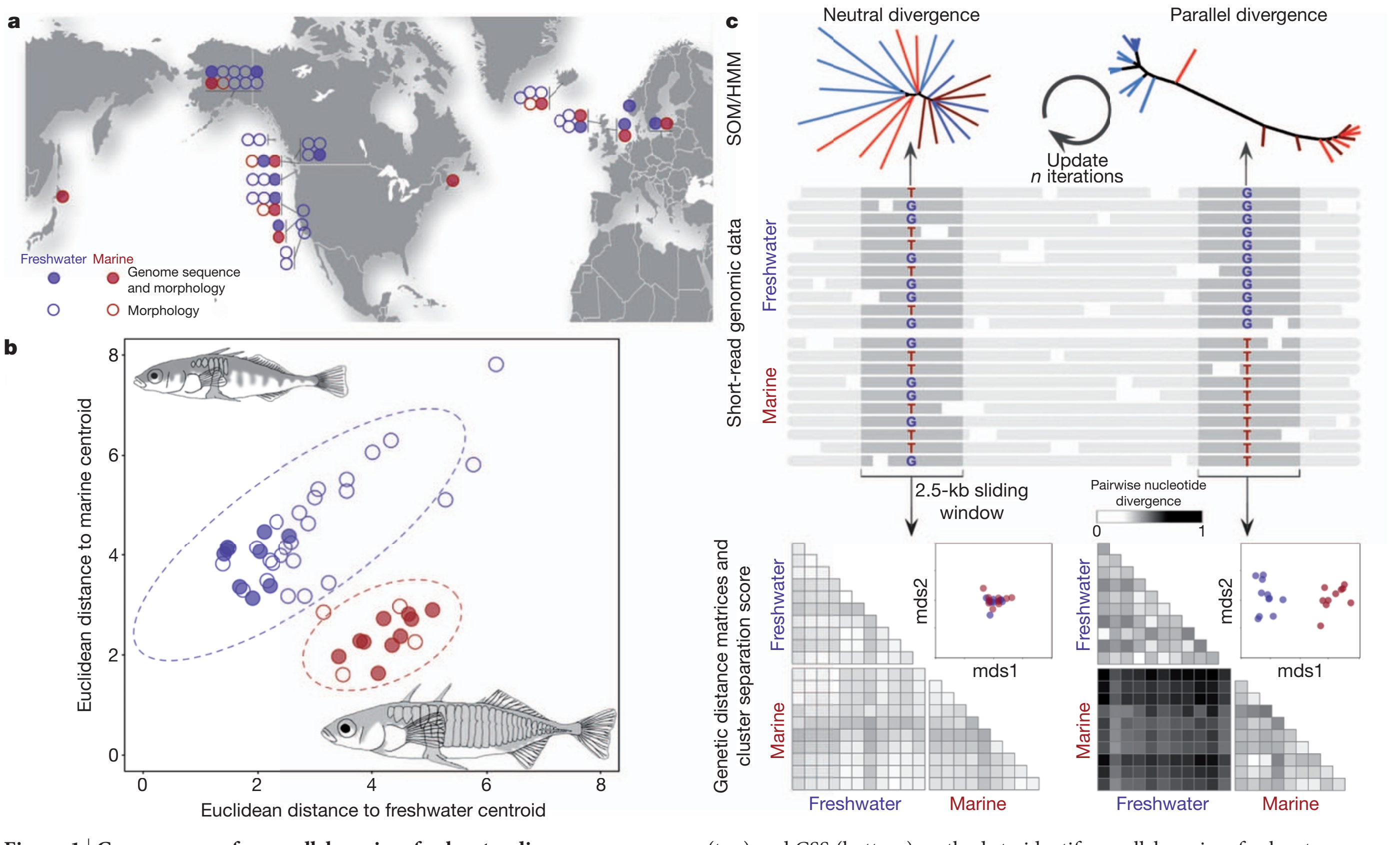

Figure 1 | Genome scans for parallel marine–freshwater divergence. **a**, Marine (red) and freshwater (blue) stickleback populations were surveyed from diverse locations. **b**, Morphometric analysis was used to select individuals for re-sequencing. The 20 chosen individuals are from multiple geographically proximate pairs of populations with typical marine and freshwater morphology (solid symbols). Points, population mean morphologies; ellipses, 95% confidence intervals for ecotypes. **c**, Genomes were analysed using SOM/HMM (top) and CSS (bottom) methods to identify parallel marine–freshwater divergent regions. Across most of the genome, the dominant patterns reflect neutral divergence or geographic structure. In contrast, <0.5% of the genome shows clustering by ecotype, a pattern characteristic of divergent marine and freshwater adaptation via parallel reuse of standing genetic variation[11,12]. Mds1 and mds2 represent the first and second major axes of variation extracted from pairwise genetic distance matrices using multidimensional scaling.

models recurring phylogenetic patterns on a local genomic basis with increasing resolution (Fig. 1c and Supplementary Information). Most of the genome was assigned to trees describing geographic relationships between populations (for example, distinct Pacific versus Atlantic clades, each containing marine and freshwater fish; Supplementary Table 7 and Supplementary Figs 2 and 3). A total of 215 regions comprising 2,096,101 bp (0.46% of the genome; median size: 4,684 bp) were assigned to one of four trees separating most marine from most freshwater fish (Supplementary Fig. 3, trees a–d). After filtering, the most prevalent marine–freshwater divergent tree identified 90 genomic regions with a median size of 4,266 bp covering 848,691 bp (0.18% of the genome).

Second, we used a genetic distance-based approach (Fig. 1c) based on building 21 × 21 pairwise nucleotide divergence (π) matrices for each of 877,568 overlapping windows across the genome (2,500 bp, step size: 500 bp). Each distance matrix was used to calculate a marine–freshwater cluster separation score (CSS), quantifying the average distance between marine and freshwater clusters after accounting for variance within ecotypes (Supplementary Information). The score is highly correlated with genetic distance (F_{ST}), but provides increased resolution under high divergence (Supplementary Fig. 4). After permutation testing, we recovered 174 marine–freshwater divergent regions, covering a total of 1,214,500 bp (0.26% of the genome; median size: 3,000 bp) at a 5% false discovery rate (FDR), and 84 divergent regions covering 479,500 bp (0.10% of the genome; median: 4,000 bp) at 2% FDR. To assign cluster membership in highly divergent genomic regions, we also used an unguided Bayesian model-based data-driven clustering (DDC; Fig. 2c and Supplementary Information). For each window of the genome, we estimated the most likely number of distinct clusters of fish ($k = 0$ to 5) and the cluster memberships.

The independent SOM/HMM and CSS approaches both successfully recover the previously described chromosome IV *EDA* locus among the top-scoring marine–freshwater divergent regions (Fig. 2). Notably, the cluster membership assigned by DDC successfully recapitulates the breakpoints of the minimal 16-kb shared freshwater *EDA* haplotype (Fig. 2c) previously defined by a multi-year positional cloning study of the major locus controlling armour plate differences in sticklebacks[11]. Additional regions were identified on the same chromosome with similar marine–freshwater divergence patterns, including regions surrounding the developmental signalling gene *WNT7B* (Supplementary Fig. 5), and a locus involved in hormone and neurotransmitter binding and metabolism (sulphotransferase 4a1, *SULT4A* (ref. 32)). SOM/HMM and CSS defined many other loci that also show globally shared marine–freshwater divergence, including 242 regions identified by either method (0.5% of the genome), and 147 regions identified by both (0.2% of the genome). The median size of recovered regions (<5 kb) approaches the size of individual genes, and often highlights purely intergenic regions, such as the exclusively non-coding region identified between *BANP* and *RAS* on chromosome XIX (Supplementary Fig. 6). The genomic distribution, sizes and overlaps of recovered regions are described in Fig. 3, Supplementary Fig. 7 and Supplementary Table 8, including a list of specific genes identified in top-scoring regions (Supplementary Data 1). Using genotyping assays for SNPs in 11 regions recovered by both SOM/HMM and CSS analyses, we found that 91% of tested regions show significant enrichment of ecotypic alleles in independent marine and freshwater populations (Supplementary Information). These results confirm that our experimental design successfully identifies both known and novel loci consistently associated with parallel evolution of distinct marine and freshwater ecotypes.

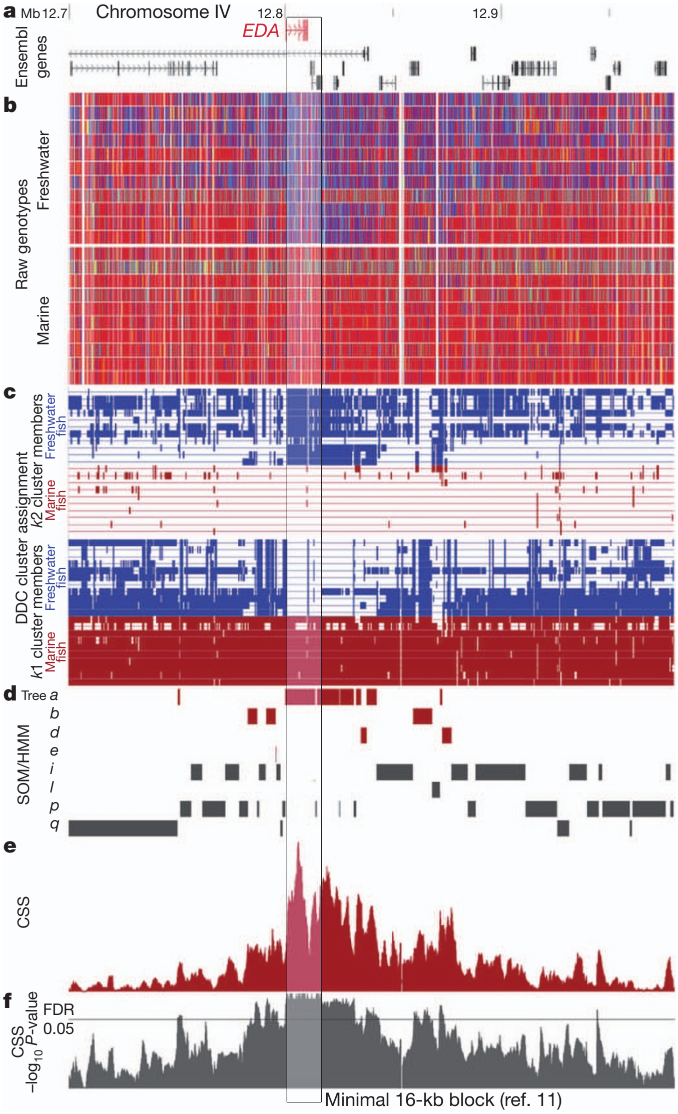

Figure 2 | Parallel divergence signals at known armour plate locus. **a**, Ensembl gene models around *EDA*. **b**, Visual genotypes for sequenced fish (homozygous sites for most frequent allele in marine fish (red); homozygous for alternative allele (blue); heterozygous (yellow), or non-variable/missing/repeat-masked data (white)). **c**, DDC cluster assignments for marine (red) and freshwater populations (blue). Most fish are assigned to cluster *k*1, except in the boxed region, where freshwater fish are assigned to a distinct cluster (*k*2). **d**, SOM/HMM analysis supports patterns of divergence with a marine–freshwater-like tree topology in the centre, but not edges, of the window (trees *a*–*d*). **e**, **f**, Similar support is shown by CSS analysis (**e**) and its associated *P*-value (**f**). The combined analyses define a consensus 16-kb region shared in freshwater fish (vertical shaded box), matching the minimal haplotype known to control repeated low armour evolution in sticklebacks[11].

Compared to the genome overall, the 242 regions implicated in repeated marine–freshwater evolution show higher gene density (Supplementary Fig. 8, $P < 4.5 \times 10^{-13}$) and higher concentration of conserved non-coding sequences in intergenic regions (Supplementary Fig. 9, $P < 1.9 \times 10^{-11}$), probably reflecting a more complex regulatory architecture[33]. Gene Ontology analysis shows significant enrichment of genes involved in cellular response to signals, behavioural interaction between organisms, amine and fatty acid metabolism, cell–cell junctions and WNT signalling (Supplementary Table 9). Changes in these biological processes, and in the individual genes defined by parallel divergence analysis, probably underlie recurrent differences in morphology, physiology and behaviour previously described in marine and freshwater sticklebacks[7]. For example, the *WNT7B* and *WNT11* family members identified by the genomic survey have previously been implicated in a paracrine signalling pathway that controls kidney collecting tubule length and diameter[34]. Fish living in fresh water produce copious hypotonic urine compared to marine fish[35], and long-term adaptation to freshwater may select for

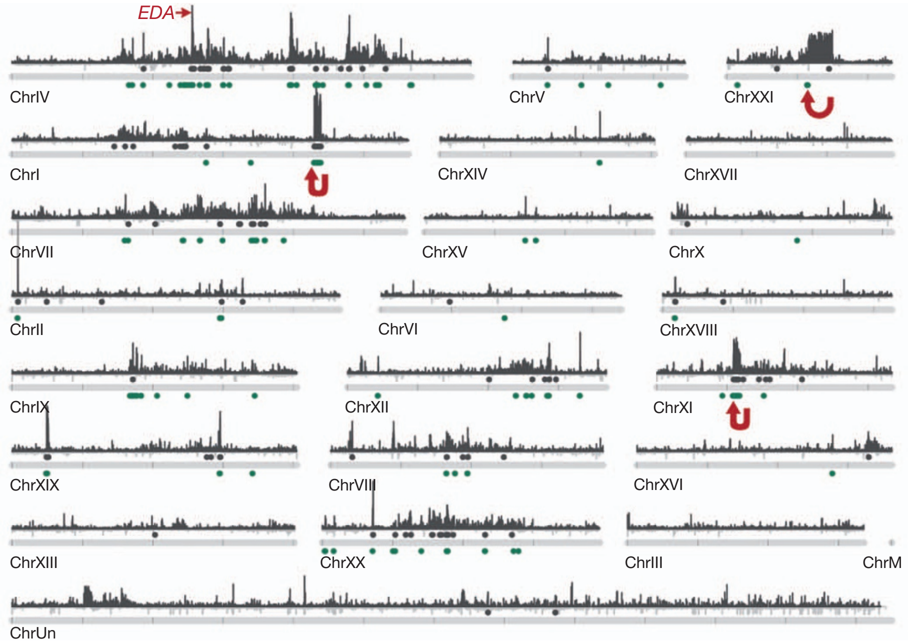

Figure 3 | Genome-wide distribution of marine–freshwater divergence regions. Whole-genome profiles of SOM/HMM and CSS analyses reveal many loci distributed on multiple chromosomes (plus unlinked scaffolds, here grouped as 'ChrUn'). Extended regions of marine–freshwater divergence on chromosomes I, XI and XXI correspond to inversions (red arrows). Marine–freshwater divergent regions detected by CSS are shown as grey peaks with grey points above chromosomes indicating regions of significant marine–freshwater divergence (FDR < 0.05). Genomic regions with marine–freshwater-like tree topologies detected by SOM/HMM are shown as green points below chromosomes.

variants in the same developmental signalling pathways that polarize epithelial cell divisions and regulate kidney tubule formation in other animals.

Extent of parallel reuse in hybrid zones

Although our method identifies regions used repeatedly during stickleback evolution, it does not tell us how prevalent such regions are among all differentiated loci in a particular marine–freshwater species pair. To address this, we analysed patterns of genomic differentiation across a marine–freshwater hybrid zone in River Tyne, Scotland (Fig. 4a). Previous studies show that ecologically mediated postzygotic selection maintains distinct ecotypes in this system, despite hybridization and opportunity for extensive gene flow[16]. Whole-genome sequencing of a pair of marine and freshwater fish from either end of the Tyne hybrid zone identified a set of genomic windows with high divergence. Within the top 0.1% divergent windows, 35.3% contain elevated globally shared marine–freshwater divergence (Fig. 4b and Supplementary Information), indicating an ancient shared origin for many, but not all, loci with highly differentiated alleles in this marine–freshwater species pair. Previous studies have shown that some traits in sticklebacks evolve by independent mutations that vary among populations[10]. The non-globally shared divergent alleles in the Tyne may also represent recent, or locally arising, adaptive variants, although further studies will be required to link such variants to particular traits, or to distinguish them from neutral but highly variable regions of the stickleback genome.

Marine–freshwater chromosome inversions

When adaptive divergence occurs in hybridizing systems, theory predicts that selection can favour molecular mechanisms that suppress recombination between independent adaptive loci[17]. We observed extended stretches of elevated CSS spanning 442 kb, 412 kb and 1,700 kb on chromosomes I, XI and XXI, respectively (Fig. 3). On the basis of sharp transitions in CSS score and DDC cluster assignments at the boundaries, we hypothesized that chromosomal inversions explain these extended regions. By analysing paired-end sequence reads from a marine large-insert (~220 kb) bacterial

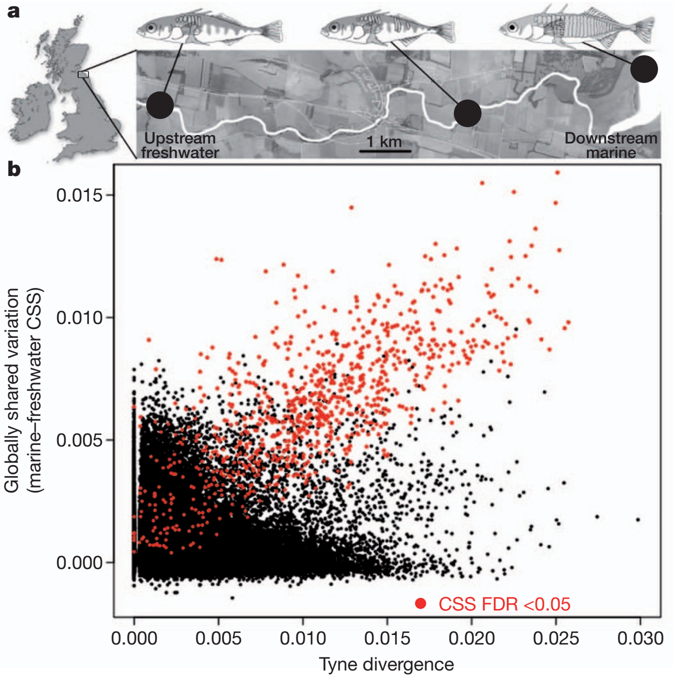

Figure 4 | How much of local marine–freshwater adaptation occurs by reuse of global variants? a, Classic marine and freshwater ecotypes are maintained in downstream and upstream locations of the River Tyne, Scotland, despite extensive hybridization at intermediate sites[16]. **b**, Pairwise sequence comparisons identify many genomic regions that show high divergence between upstream and downstream fish (*x* axis). Many, but not all, of these regions also show high global marine–freshwater divergence (*y* axis; red points indicate significant CSS FDR < 0.05), indicating that both global and local variants contribute to formation and reproductive isolation of a marine–freshwater species pair.

artificial chromosome (BAC) library[36], we identified individual clones with size and orientation anomalies relative to the freshwater reference genome assembly. The only locations with five or more anomalous clones mapped to chromosomes I, XI and XXI, and these anomalies could be resolved by the presence of inverted chromosome segments between the marine fish and the freshwater reference genome (Fig. 5a, b). Sequences flanking the predicted inversion breakpoints contain inverted repeats, consistent with generation of inversions by intra-chromosomal recombination (Supplementary Fig. 10). Notably, repeats flanking the chromosome XI inversion contained alternative 3′ exons for the voltage-gated potassium channel gene *KCNH4*. Because *KCNH4* transcription is initiated within the inversion, alternative inversion orientations could generate marine- and freshwater-specific KCNH4 isoforms (Fig. 5c). Although the functional consequences of such ecotype-specific isoforms remain unknown, KCNH4 homologues help to maintain resting currents, affect cardiac contractility, and alter performance on cognitive tasks if perturbed in mice[37–39]. Furthermore, QTL for two distinct marine–freshwater divergent traits have previously been mapped to the broad region of the chromosome XXI inversion (Fig. 5d)[27,30], as expected if inversions help to maintain linkage between different adaptive QTLs[40].

Importantly, cluster assignment of individual fish by DDC shows that most marine and freshwater populations in the Pacific carry contrasting forms of the inversion regions (Supplementary Table 10). Similar ecotype associations are seen in the Atlantic basin for chromosome I (no exceptions), XI (two exceptions), and to a lesser extent for chromosome XXI (three freshwater exceptions). Genetic markers within the chromosome I and XXI regions are polymorphic in hybrid zones, and show large frequency differences when genotyped in adjacent upstream and downstream fish, confirming that these regions are subject to divergent selection in marine and freshwater habitats (Supplementary Table 10). Our results help to explain the broader patterns of genomic divergence seen in Fig. 3, and add to growing evidence that chromosome inversions are a common genomic mechanism that maintains contrasting ecotypes in hybridizing natural populations[41–44].

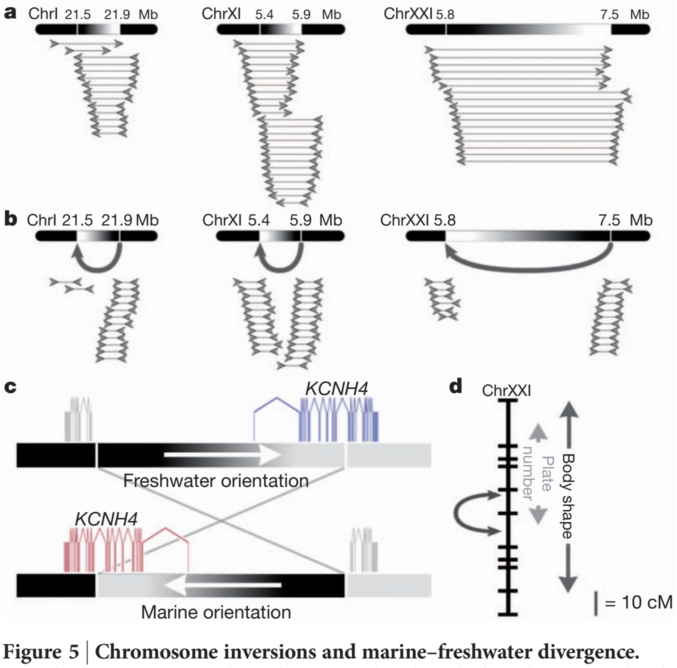

Figure 5 | Chromosome inversions and marine–freshwater divergence. **a**, Multiple marine BAC clones have paired-end reads that place anomalously against the freshwater reference genome (grey arrows below chromosome bars; see Supplementary Methods for BAC names). **b**, Intrachromosomal inversions on chromosomes I, XI and XXI resolve orientation and size anomalies for all marine clones. **c**, The chromosome XI inversion breakpoints map inside the exons of *KCNH4*, a potassium transporter gene. Duplicated 3′ exons lead to different transcript orientations and gene products in marine (red gene model) and freshwater fish (blue gene model). **d**, The chromosome XXI inversion occurs in a region with separate QTLs controlling armour plate number and body shape[11,30], traits that differ between marine and freshwater fish.

Proportion of regulatory and coding change

Identification of a genome-wide set of loci used repeatedly in stickleback adaptation provides a rare opportunity to estimate the relative contribution of coding and regulatory changes underlying adaptive evolution in natural populations. To examine this issue, we analysed 64 marine–freshwater divergent regions with the strongest evidence of parallel evolution: those identified by both SOM/HMM and CSS analyses using the strictest significance thresholds (Supplementary Information and Supplementary Data 1), and containing SNPs showing perfect allele–ecotype association between marine and freshwater fish. Many of these 64 regions (41%) mapped entirely to non-coding regions of the genome, and presumably contain regulatory changes (Fig. 6a). A smaller fraction contains protein-coding sequences with consistent non-synonymous substitutions between marine and freshwater fish (17%). Finally, a fraction of regions (43%) include both coding and non-coding sequences (including non-coding RNAs), but lack ecotype-specific amino acid substitutions (Supplementary Data 1). Because all of these regions contain SNPs with perfect allele–ecotype association that do not cause protein-coding changes, they also probably contribute to adaptive divergence by regulatory alterations. The combined data suggest that both coding and regulatory differences contribute to parallel stickleback evolution, with regulatory changes accounting for a much larger proportion of the overall set of loci repeatedly selected during marine–freshwater divergence.

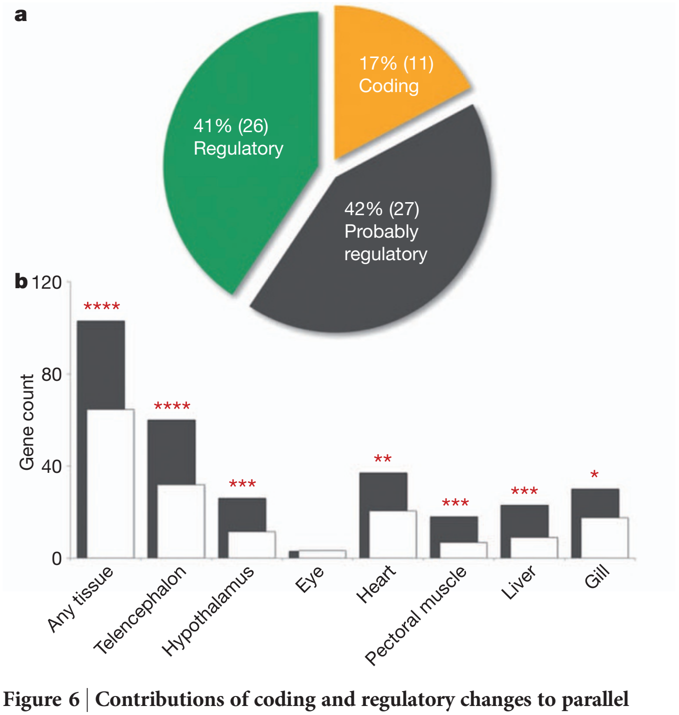

Figure 6 | Contributions of coding and regulatory changes to parallel marine–freshwater stickleback adaptation. **a**, A genome-wide set of marine–freshwater divergent loci recovered by both SOM/HMM and CSS analyses includes regions with consistent amino acid substitutions between marine and freshwater ecotypes (yellow sector); regions with no predicted coding sequence (green sector); and regions with both coding and non-coding sequences, but no consistent marine–freshwater amino acid substitutions (grey). **b**, Genome-wide expression analysis shows that marine–freshwater regions identified by SOM/HMM or CSS analyses are enriched for genes showing significant expression differences in 6 out of 7 tissues between marine LITC and freshwater FTC fish (observed, grey bars; expected, white bars; *$P < 0.01$, **$P < 0.001$, ***$P < 0.0001$, ****$P \ll 0.00001$), consistent with a role for regulatory changes in marine–freshwater evolution.

To assess further the possible role of gene regulatory evolution in stickleback evolution, we constructed whole-genome expression arrays to compare levels of gene expression in tissues from Little Campbell River (LITC) marine and Fish Trap Creek (FTC) freshwater fish. Of 12,594 informative genes across the genome, 2,817 showed significant expression differences between ecotypes. Genes with marine–freshwater expression differences were significantly more likely to occur in or near the adaptive regions recovered by SOM/HMM or CSS analysis (Fig. 6b, $P < 7.1 \times 10^{-8}$). Although expression differences can be due to either *cis*- or *trans*-acting changes, the expression data are consistent with an important role of regulatory changes during parallel evolution of marine and freshwater sticklebacks.

Discussion

Progress in genetic mapping and positional cloning approaches has recently made it possible to identify a few individual genes and mutations that contribute to phenotypic differences between stickleback populations[10–12,25]. Despite this progress, identifying many such examples using genetic linkage mapping alone would require years of additional effort. Fortunately, the highly replicated nature of stickleback evolution provides clear molecular signatures that can be used to recover many loci consistently associated with parallel marine–freshwater adaptation. The signal resolution of repeatedly used adaptive loci approaches ~5 kb, often identifying single genes or intergenic regions, and offering a significant advantage over the several hundred kilobase candidate intervals typically identified in genetic mapping crosses[11,12], or the megabase or larger regions identified in previous selection scans of the stickleback genome[13]. The many marine–freshwater divergent loci and gene expression changes identified in the current study will substantially accelerate ongoing searches for the genetic and molecular basis of fitness-related morphological, physiological and behavioural differences between marine and freshwater fish.

In addition, the genome-wide set of divergent regions already provides new insights into evolutionary processes shaping adaptive evolution and ecological speciation. Our results indicate that parallel evolution of marine and freshwater sticklebacks occurs by dynamic reassembly of many 'islands' of divergence distributed across many chromosomes. Reassembly by linkage is probably strengthened by inversions that distinguish marine and freshwater ecotypes. Differences in both globally shared and locally restricted genetic variation actively maintained across a hybrid zone provide a snapshot of the genomic architecture and evolutionary processes contributing to the early stages of reproductive isolation. Finally, our data indicate that repeated evolution of marine–freshwater differences depends on both protein-coding and regulatory changes. Regulatory evolution seems to have a particularly prominent role, as indicated by the increased density of conserved non-coding intergenic sequences found near marine–freshwater divergent loci (Supplementary Fig. 9); the substantial fraction of loci mapping entirely to non-coding regions (Fig. 6a); and the significant enrichment of genes with expression differences near key regions used for parallel evolution (Fig. 6b). Mutations causing structural changes in proteins are the most abundant variants recovered in laboratory *Escherichia coli* and yeast evolution experiments[45,46]. They make up 90% of 40 published examples of adaptive changes between closely related taxa[4], and 63–77% of the known molecular basis of phenotypic traits in domesticated or wild species[5]. The larger fraction of regulatory changes implicated during repeated stickleback evolution may reflect our use of whole-genome rather than candidate gene approaches, stronger selection against loss-of-function and pleiotropic protein-coding changes in natural populations than in laboratory or domesticated organisms[1–3], or an increasing prevalence of regulatory changes at interspecific compared to intraspecific levels[5,47], including emerging species such as marine and freshwater sticklebacks.

Although our study has focused on marine–freshwater divergence, freshwater sticklebacks also repeatedly evolve characteristic lake–stream differences; open-water and bottom-dwelling lake ecotypes; gigantism in particular lakes; and substantial changes in seasonality and life history[6,7,48–50]. Given the considerable fraction of parallel stickleback evolution probably occurring by shared variants (Fig. 4b), sequencing of additional populations should make it possible to identify similarly shared loci contributing to other ecological traits, again using the power of replicated evolution to illuminate both specific and general mechanisms underlying evolutionary change in natural populations.

METHODS SUMMARY

A reference stickleback genome sequence was assembled from a single female freshwater stickleback (Bear Paw Lake, Alaska), using 9× coverage of paired-end Sanger-sequenced reads from multiple insert size libraries. Scaffolds were assigned to linkage groups in a genetic cross, and annotation was carried out using the Ensembl evidence-based pipeline. Twenty-one fish from independent populations were chosen for short-read sequencing (48× combined coverage) based on morphometric analysis. Patterns of genetic variation were analysed for divergence between marine and freshwater fish, using both a self-organizing map/Hidden Markov Model and a pairwise distance matrix approach (see Supplementary Information). Paired-end reads from a marine BAC library were placed against the reference freshwater genome sequence to identify possible chromosome rearrangements. Sequenom iPlex genotyping assays were carried out to verify predicted SNPs and divergent marine–freshwater regions. RNA samples were prepared from tissues of marine and freshwater fish born and raised under identical laboratory conditions. Significant expression differences were detected with Agilent microarrays using eBayes (limma R package). GO category enrichments were analysed using GOstats (BioConductor 2.7). Additional methods and analyses are provided in Supplementary Information.

Received 17 September 2011; accepted 8 February 2012.

1. Stern, D. L. Perspective: Evolutionary developmental biology and the problem of variation. *Evolution* **54,** 1079–1091 (2000).
2. Carroll, S. B. Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. *Cell* **134,** 25–36 (2008).
3. Wray, G. The evolutionary significance of *cis*-regulatory mutations. *Nature Rev. Genet.* **8,** 206–216 (2007).
4. Hoekstra, H. E. & Coyne, J. A. The locus of evolution: evo devo and the genetics of adaptation. *Evolution* **61,** 995–1016 (2007).
5. Stern, D. L. & Orgogozo, V. The loci of evolution: how predictable is genetic evolution? *Evolution* **62,** 2155–2177 (2008).
6. McKinnon, J. S. & Rundle, H. D. Speciation in nature: the threespine stickleback model systems. *Trends Ecol. Evol.* **17,** 480–488 (2002).
7. Bell, M. A. & Foster, S. A. *The Evolutionary Biology of the Threespine Stickleback* (Oxford Univ. Press, 1994).
8. Endler, J. A. Natural selection in the wild. *Monogr. Popul. Biol.* **21,** 1–336 (1986).
9. Kingsley, D. M. & Peichel, C. L. The molecular genetics of evolutionary change in sticklebacks. In *Biology of the Threespine Stickleback* 41–81 (CRC Press, 2007).
10. Chan, Y. F. *et al.* Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a Pitx1 enhancer. *Science* **327,** 302–305 (2010).
11. Colosimo, P. F. *et al.* Widespread parallel evolution in sticklebacks by repeated fixation of Ectodysplasin alleles. *Science* **307,** 1928–1933 (2005).
12. Miller, C. T. *et al.* cis-Regulatory changes in *Kit ligand* expression and parallel evolution of pigmentation in sticklebacks and humans. *Cell* **131,** 1179–1189 (2007).
13. Hohenlohe, P. A. *et al.* Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLoS Genet.* **6,** e1000862 (2010).
14. Kitano, J. *et al.* Adaptive divergence in the thyroid hormone signaling pathway in the stickleback radiation. *Curr. Biol.* **20,** 2124–2130 (2010).
15. Hagen, D. Isolating mechanisms in threespine sticklebacks (*Gasterosteus*). *J. Fish. Res. Board Can.* **24,** 1637–1692 (1967).
16. Jones, F., Brown, C., Pemberton, J. & Braithwaite, V. Reproductive isolation in a threespine stickleback hybrid zone. *J. Evol. Biol.* **19,** 1531–1544 (2006).
17. Barton, N. H. & Gale, K. S. Genetic analysis of hybrid zones. In *Hybrid Zones and the Evolutionary Process* 13–45 (Oxford Univ. Press, 1993).
18. Vinogradov, A. E. Genome size and GC-percent in vertebrates as determined by flow cytometry: the triangular relationship. *Cytometry* **31,** 100–109 (1998).
19. Aparicio, S. *et al.* Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. *Science* **297,** 1301–1310 (2002).
20. Jaillon, O. *et al.* Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. *Nature* **431,** 946–957 (2004).
21. Kasahara, M. *et al.* The medaka draft genome and insights into vertebrate genome evolution. *Nature* **447,** 714–719 (2007).
22. Star, B. *et al.* The genome sequence of Atlantic cod reveals a unique immune system. *Nature* **477,** 207–210 (2011).
23. Kawahara, R. & Nishida, M. Extensive lineage-specific gene duplication and evolution of the spiggin multi-gene family in stickleback. *BMC Evol. Biol.* **7,** 209 (2007).
24. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15,** 1034–1050 (2005).

25. Shapiro, M. D. *et al.* Genetic and developmental basis of evolutionary pelvic reduction in threespine sticklebacks. *Nature* **428,** 717–723 (2004).
26. Peichel, C. L. *et al.* The genetic architecture of divergence between threespine stickleback species. *Nature* **414,** 901–905 (2001).
27. Colosimo, P. F. *et al.* The genetic architecture of parallel armor plate reduction in threespine sticklebacks. *PLoS Biol.* **2,** 635–641 (2004).
28. Cresko, W. A. *et al.* Parallel genetic basis for repeated evolution of armor loss in Alaskan threespine stickleback populations. *Proc. Natl Acad. Sci. USA* **101,** 6050–6055 (2004).
29. Kimmel, C. B. *et al.* Evolution and development of facial bone morphology in threespine sticklebacks. *Proc. Natl Acad. Sci. USA* **102,** 5791–5796 (2005).
30. Albert, A. Y. K. *et al.* The genetics of adaptive shape shift in stickleback: pleiotropy and effect size. *Evolution* **62,** 76–85 (2008).
31. Barrett, R. D. H., Rogers, S. M. & Schluter, D. Natural selection on a major armor gene in threespine stickleback. *Science* **322,** 255–257 (2008).
32. Allali-Hassani, A. *et al.* Structural and chemical profiling of the human cytosolic sulfotransferases. *PLoS Biol.* **5,** e97 (2007).
33. Knecht, A. K., Hosemann, K. E. & Kingsley, D. M. Constraints on utilization of the EDA-signaling pathway in threespine stickleback evolution. *Evol. Dev.* **9,** 141–154 (2007).
34. Yu, J. *et al.* A *Wnt7b*-dependent pathway regulates the orientation of epithelial cell division and establishes the cortico-medullary axis of the mammalian kidney. *Development* **136,** 161–171 (2009).
35. Marshall, W. S. & Grosell, M. Ion transport, osmoregulation and acid-base balance. In *The Physiology of Fishes* 177–230 (CRC Press, 2006).
36. Kingsley, D. M. *et al.* New genomic tools for molecular studies of evolutionary change in threespine sticklebacks. *Behaviour* **141,** 1331–1344 (2004).
37. Miyake, A., Mochizuki, S., Yokoi, H., Kohda, M. & Furuichi, K. New ether-à-go-go K^+ channel family members localized in human telencephalon. *J. Biol. Chem.* **274,** 25018–25025 (1999).
38. Miyake, A. *et al.* Disruption of the ether-à-go-go K^+ channel gene BEC1/KCNH3 enhances cognitive function. *J. Neurosci.* **29,** 14637–14645 (2009).
39. Gutman, G. A. *et al.* International Union of Pharmacology. LIII. Nomenclature and molecular relationships of voltage-gated potassium channels. *Pharmacol. Rev.* **57,** 473–508 (2005).
40. Kirkpatrick, M. & Barton, N. Chromosome inversions, local adaptation and speciation. *Genetics* **173,** 419–434 (2006).
41. Hoffmann, A. A. & Rieseberg, L. H. Revisiting the impact of inversions in evolution: from population genetic markers to drivers of adaptive shifts and speciation? *Annu. Rev. Ecol. Evol. Syst.* **39,** 21–42 (2008).
42. Lowry, D. B. & Willis, J. H. A widespread chromosomal inversion polymorphism contributes to a major life-history transition, local adaptation, and reproductive isolation. *PLoS Biol.* **8,** e1000500 (2010).
43. Joron, M. *et al.* Chromosomal rearrangements maintain a polymorphic supergene controlling butterfly mimicry. *Nature* **477,** 203–206 (2011).
44. Feder, J. L., Roethele, J. B., Filchak, K., Niedbalski, J. & Romero-Severson, J. Evidence for inversion polymorphism related to sympatric host race formation in the apple maggot fly, *Rhagoletis pomonella*. *Genetics* **163,** 939–953 (2003).
45. Barrick, J. E. *et al.* Genome evolution and adaptation in a long-term experiment with *Escherichia coli. Nature* **461,** 1243–1247 (2009).
46. Kvitek, D. J. & Sherlock, G. Reciprocal sign epistasis between frequently experimentally evolved adaptive mutations causes a rugged fitness landscape. *PLoS Genet.* **7,** e1002056 (2011).
47. Wittkopp, P. & Haerum, B. K. Regulatory changes underlying expression differences within and between *Drosophila* species. *Nature Genet.* **40,** 346–350 (2008).
48. Reimchen, T. E., Stinson, E. M. & Nelson, J. S. Multivariate differentiation of parapatric and allopatric populations of threespine stickleback in the Sangan River watershed, Queen Charlotte Islands. *Can. J. Zool.* **63,** 2944–2951 (1985).
49. Deagle, B. E. *et al.* Population genomics of parallel phenotypic evolution in stickleback across stream–lake ecological transitions. *Proc. R. Soc. B* **279,** 1277–1286 (2011).
50. McPhail, J. D. Speciation and the evolution of reproductive isolation in the sticklebacks (*Gasterosteus*) of south-western British Columbia. In *The Evolutionary Biology of the Threespine Stickleback* 399–437 (Oxford Univ. Pres, 1994).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements Stickleback sequencing at Broad Institute was supported by grants from the National Human Genome Research Institute (NHGRI). R.M.M., J.S., J.G. and D.M.K. were supported by NHGRI CEGS Grant P50-HG002568; Y.F.C. by a Stanford Affymetrix Bio-X Graduate Fellowship; C.T.M. by the Jane Coffins Childs Fund; and B.R.S., T.R.H. and A.A.P. by graduate fellowships from NSF and NDSEG. D.M.K. is an investigator of the Howard Hughes Medical Institute. K.L.-T. is a EURYI award recipient funded by ESF. We thank W. Cresko for the BEPA individual used for reference genome sequencing; M. Bell, J. McKinnon, B. Jónsson, S. Mori, C. Peichel, D. Schluter, M. Kalbe, T. Reimchen, D.-P. Højgaard, M. McLaughlin, B. Geyti and B. Blackman for discussions and assistance with specimens used in population surveys; and G. Bejerano for useful discussions and assistance with computational analysis.

Author Contributions K.L.-T., F.D.P., E.S.L. and D.M.K. planned and oversaw the project and K.L.-T. and D.M.K. are co-senior authors. E.M. and M.G.G. assembled, J.S., J.G., M.C.D., A.K.K. and R.M.M. anchored, and S.W., E.B. and S.S. annotated the reference genome. C.A. constructed the BEPA BAC library. F.C.J., Y.F.C., D.M.K., K.L.-T., F.D. and M.G.G. designed the whole-genome re-sequencing experiment. F.C.J. and Y.F.C. performed morphometric analyses. S.D.B. and J.J. prepared and coordinated samples. M.G.G., M.P. and M.C.Z. analysed pilot data and performed simulations to evaluate sequencing strategies. M.G.G., P.R., E.M., F.C.J., Y.F.C., J.J. and R.S. analysed polymorphisms. P.R. and M.G.G. developed and carried out the SOM/HMM analysis. F.C.J. and Y.F.C. developed and carried out the CSS and DDC analysis. P.R. analysed gene and non-coding element density, and performed phylogenetic analysis. T.H. analysed GO term enrichments. F.C.J. and Y.F.C. carried out hybrid zone analysis. C.T.M., B.R.S., J.G., J.S., Y.F.C. and F.C.J. analysed chromosome inversions. F.C.J. and D.M.K. performed analysis of coding and regulatory changes. H.Z., A.A.P. and T.H. performed the whole-genome expression study. D.M.K., F.C.J., Y.F.C., P.R., F.D.P. and K.L.-T. wrote the paper with input from other authors.

Author Information UCSC Genome browser tracks showing genome-wide analyses are available at http://sticklebrowser.stanford.edu. Microarray expression data are deposited at the Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo) under accession number GSE34783. BAC-end sequences are deposited at http://www.ncbi.nlm.nih.gov/dbGSS (accession numbers JS583469 to JS583576). Reprints and permissions information is available at www.nature.com/reprints. This paper is distributed under the terms of the Creative Commons Attribution-Non-Commercial-Share Alike licence, and is freely available to all readers at www.nature.com/nature. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to K.L.-T. (kersli@broad.mit.edu) or D.M.K. (kingsley@stanford.edu).

Broad Institute Genome Sequencing Platform & Whole Genome Assembly Team

Jen Baldwin[1], Toby Bloom[1], David B. Jaffe[1], Robert Nicol[1] & Jane Wilkinson[1]

[1]Broad Institute of MIT and Harvard, 7 Cambridge Center, Cambridge, Massachusetts 02142, USA.