Predicting future from past: The genomic basis of recurrent and rapid stickleback evolution


Similar forms often evolve repeatedly in nature, raising long-standing questions about the underlying mechanisms. Here, we use repeated evolution in stickleback to identify a large set of genomic loci that change recurrently during colonization of freshwater habitats by marine fish. The same loci used repeatedly in extant populations also show rapid allele frequency changes when new freshwater populations are experimentally established from marine ancestors. Marked genotypic and phenotypic changes arise within 5 years, facilitated by standing genetic variation and linkage between adaptive regions. Both the speed and location of changes can be predicted using empirical observations of recurrenc in natural populations or fundamental genomic features like allelic age, recombination rates, density of divergent loci, and overlap with mapped traits. A composite model trained on these stickleback features can also predict the location of key evolutionary loci in Darwin’s finches, suggesting that similar features are important for evolution across diverse taxa.

INTRODUCTION

Can evolutionary outcomes be predicted? Biologists have long been fascinated with this question, including Darwin and Wallace’s anticipation of the existence of Morgan’s sphinx moth based on orchid morphology (1, 2). Vavilov’s prediction of the types of morphological variants likely to occur in plants (3), and Gould’s gedankenexperiment about replaying the tape of life (4). Natural examples of recurrent evolution provide a particularly favorable opportunity to study the mechanisms that influence evolutionary predictability, including molecular patterns (5, 6).

Although the predictability of evolution may appear to be in conflict with the unpredictability of historical contingency, understanding the past can yield important insights into future evolution. For example, vertebrate populations frequently harbor large reservoirs of standing genetic variation (SGV) (7) that give independent populations access to similar raw genetic material to respond to environmental challenges, as observed in diverse species including songbirds, cichlid fishes, and the threespine stickleback (Gasterosteus aculeatus) (8–11). SGV is often apparent in divergent species or populations where it is pretested by natural selection and then distributed by hybridization to related populations. Thus filtered and capable of leaping up fitness landscapes, SGV can also drive rapid evolution (12), helping address a very real practical challenge to testing evolutionary predictions: time.

Longitudinal studies of evolving populations have been used to estimate the tempo and strength of selection on a variety of traits in different species (13–18). Rapid phenotypic evolution over contemporary time scales has enabled hypothesis testing against detailed observations at every step in the process. There is an increasing and impressive body of research examining the genomic consequences of these phenotypic changes in microbial, invertebrate, and vertebrate systems (19–26).

Stickleback fish provide an outstanding system for further study of the genomic basis of recurrent evolution. At the end of the last Ice Age, threespine stickleback, including anadromous populations that migrate from the ocean to freshwater environments to breed, colonized and adapted to countless newly exposed freshwater environments created in the wake of retreating glaciers around the northern hemisphere (27, 28). This massively parallel adaptive radiation was facilitated by natural selection acting on extensive ancient SGV (8, 11). Under the “transporter” hypothesis, these variants are maintained at low frequencies in the marine populations by low levels of gene flow from freshwater populations (29). Reuse of ancient standing variants has enabled identification of genomewide sets of loci that are repeatedly differentiated among long-established stickleback populations (8, 30–35). In addition, SGV enables new freshwater stickleback populations to evolve markedly within decades (17, 36–38), including conspicuous phenotypic changes in armor plates (17) and body shape (39).

The rapidity of stickleback evolution has made it possible to begin characterizing genomic and allele frequency changes seen in very young or newly established populations under intense directional selection on multiple traits (18, 36–38, 40–43). Here, we identify key molecular features that underlie repeated and rapid evolution of freshwater stickleback by comparing genomes from diverse extant populations with the earliest generation-by-generation changes in a detailed genomic time series from three newly founded populations. We identify several basic genomic and genetic features that
RESULTS

Global resequencing and EcoPeak identification

Previous whole-genome sequencing (WGS) of threespine stickleback identified 174 loci covering 1.2 Mb with alleles shared by common descent repeatedly selected in freshwater populations around the world (8). Just as human genetic diversity is greatest in Africa, where Homo sapiens arose (44), we hypothesized that the north Pacific region where stickleback originated (27) may contain a particularly rich pool of ancient adaptive alleles. To test this hypothesis, we generated whole-genome sequence data with 76–base pair (bp) paired-end Illumina reads for 38 new marine and 110 new freshwater stickleback, respectively (mean coverage of 5.5x) (sections S2, S4, S6, and S7). Combined with previous stickleback sequencing (8, 41), our dataset includes 227 individual genomes: 135 genomes from 70 north-east Pacific populations in Alaska, Haida Gwaii, British Columbia, and Washington and 92 genomes from 62 populations in California, Japan, and the Atlantic coasts of North America, Iceland, and northern Europe (Fig. 1A and section S8).

We used two methods to identify loci repeatedly differentiated in freshwater populations, both based on the expectation that variants recurrently selected from SGV will be more similar among geographically separated freshwater populations than neutral loci (section S9). First, we used a genetic distance–based approach within overlapping 2500-bp windows tiled across the genome [as in the study by Jones et al. (8)]. While statistically powerful, this approach may miss younger loci with few differences between alleles and exhibits spatial resolution dependent on window size. Second, we analyzed the distribution of variants at individual bases across the genome, which has base pair–level resolution and less bias against younger loci, thus at the cost of statistical power. After calling P–value–based peaks of ecotypic (freshwater–marine–associated) differentiation using both methods, we accepted calls at two stringency levels, either requiring agreement between the two analyses at 1% false discovery rate (FDR) (specific) or support from either at 5% FDR (sensitive). We refer to these peaks of ecotypic differentiation as EcoPeaks. We called EcoPeaks for different geographic sets of samples to find alleles that were either shared globally, within the northeast Pacific, or within other geographic regions.

Although results of the global analysis largely matched a previous report [79 of 81 most stringent calls from Jones et al. (8) in sensitive EcoPeaks (P = 4.2 × 10−21; table S3)], both the sensitive and specific call sets identified approximately five times as many Pacific EcoPeaks as global EcoPeaks, spanning sevenfold more of the genome (Fig. 1, E and F, and Table 1). In addition, many northeast Pacific EcoPeaks not overlapping the globally shared regions identified by Jones et al. (8) exhibit even more consistent ecotypic differentiation (assessed by P values) than others shared around the world (Fig. 1, B and C). Much smaller sets of non-global EcoPeaks were identified in the North Atlantic, subglacial Pacific, and supraglacial geographic regions (fig. S5), consistent with other reports (8, 35).

As theoretical studies indicate that SGV is immediately available for evolution and may show an increased likelihood of large–effect alleles being advantageous compared to de novo mutations (12, 45), the rich genetic reservoir observed in the northeast Pacific provides a favorable system for studying the dynamics and predictability of rapid evolutionary change (section S10). Previous studies suggest that stickleback in the northeast Pacific can adapt to freshwater environments within decades (36). However, thus far, studies have lacked temporal resolution of genome evolution in the critical early years of adaptation.

Rapid contemporary evolution and TempoPeak identification

To characterize the earliest stages of evolution after the establishment of new freshwater populations, we analyzed annual samples from populations that were recently founded by anadromous stickleback in three lakes in Alaska (Fig. 2A and section S1). In 1982, stickleback in Loberg Lake (LB) were exterminated to improve recreational fishing (17). Sometime between 1983 and 1988, LB was invaded by completely plated (~33 plates per side) anadromous stickleback [most likely from neighboring Rabbit Slough (RS)]. The characteristic freshwater, armor-reduced phenotype increased rapidly from ~16% in 1991 to ~50% by 1995 and to ~95% by 2017 (Fig. 2B) (17), with similarly rapid changes in overall body shape (39) and reproductive patterns (46). So as to more systematically examine even earlier generations of freshwater adaptation, Bell et al. (47) introduced ~3000 anadromous RS fish into each of two other Cook Inlet lakes without outlets that had been similarly treated to exterminate fish: Cheney Lake (CH) in 2009 and Scout Lake (SC) in 2011. Low–armor-plated (~5 to 7 plates per side) stickleback began to appear in the second and third generation after founding in CH and SC respectively, and, by 2017, they had increased to 20 to 30% (Fig. 2B).

To obtain genomewide allele frequencies across our time series, we performed pooled WGS (pool–seq) on all seven available annual samples from CH and SC since founding and eight from LB distributed between 1999 and 2017 (Fig. 2A and sections S3, S4, S7, and S13). Each freshwater pool–seq experiment consisted of 100 individuals (with three exceptions), with mean coverage of 223× per pool. In addition, we resequenced a pool of 200 anadromous RS individuals used to found the CH population in 2009 (RS2009) to 585×.

We identified single-nucleotide polymorphisms (SNPs) with significant allele frequency changes, indicating directional selection, using a modified Cochran–Mantel–Haenszel (CMH) test optimized for pool–seq data (48), followed by an approach analogous to our EcoPeak analysis to define both a permissive “sensitive” and a stringent “specific” set of loci that we term TempoPeaks (sections S16 to S18). Combining all three populations into a single CMH analysis (CH + SC + LB) and using RS2009 as a proxy for the founders of LB, we identified 524 sensitive and 344 specific TempoPeaks. Despite operating over very different time spans, the visual correspondence between the Pacific EcoPeaks in long–established populations and the TempoPeaks in recently established populations is notable, particularly for the specific TempoPeaks, of which 323 of 344 (94%) overlap with the sensitive Pacific EcoPeaks (Fig. 2D and section S18). In contrast, even the most lenient set of global EcoPeaks and regions from Jones et al. (8) overlap only 96 of 344 (28%) and 47 of 344 (14%) specific TempoPeaks, respectively (tables S9 and S10), emphasizing the importance of understanding the locally available SGV. Even analyzing only CH + SC (thus focusing on <10 years of freshwater adaptation), we identified 271 sensitive and 86 specific TempoPeaks, 73% and 99% of which, respectively, overlap the sensitive Pacific EcoPeaks. This marked congruency suggests that the ancient SGV represented by Pacific EcoPeaks is the primary
Table 1. Overview of EcoPeaks and TempoPeaks. The comparisons by Jones et al. (8) are with the cluster separation score 5% FDR set (8).

<table>
<thead>
<tr>
<th></th>
<th>Global EcoPeaks (specific)</th>
<th>Pacific EcoPeaks (specific)</th>
<th>Pacific EcoPeaks (sensitive)</th>
<th>Cheney + Scout</th>
<th>All young</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of regions</td>
<td>39</td>
<td>209</td>
<td>212</td>
<td>86</td>
<td>344</td>
</tr>
<tr>
<td>Total bases (%)</td>
<td>3.7 Mb (0.78%)</td>
<td>27.4 Mb (5.82%)</td>
<td>91.9 Mb (19.53%)</td>
<td>3.3 Mb (6.95%)</td>
<td>17.57 Mb (3.73%)</td>
</tr>
<tr>
<td>Median size</td>
<td>21.4 kb</td>
<td>80.2 kb</td>
<td>122.9 kb</td>
<td>21.7 kb</td>
<td>27.3 kb</td>
</tr>
<tr>
<td>Recovery of regions identified by Jones et al. (8)</td>
<td>86/174 (49.4%)</td>
<td>112/174 (64.3%)</td>
<td>158/174 (90.8%)</td>
<td>47/174 (27.0%)</td>
<td>98/174 (56.3%)</td>
</tr>
<tr>
<td>Fraction in regions identified by Jones et al. (8)</td>
<td>18/39 (46.2%)</td>
<td>29/209 (13.9%)</td>
<td>33/212 (15.6%)</td>
<td>10/86 (11.6%)</td>
<td>24/344 (7.0%)</td>
</tr>
</tbody>
</table>

Fig. 1. Recurrent peaks of ecological sequence differentiation between marine and freshwater stickleback from different regions of world. (A) Marine (red) and freshwater (blue) stickleback from the locations shown were used for various analyses (table S2). (B) Detail of part of chrIV for single-nucleotide polymorphism (SNP)-based analysis of differential allele distribution between marine and freshwater ecotypes in the northeast Pacific basin. SNPs within specific-threshold EcoPeaks are red. A subset of regions overlap the globally shared peaks of marine-freshwater differentiation indicated by blue-colored bars [cluster separation score (CSS), 5% false discovery rate (FDR) identified by Jones et al. (8)]. (C) As in (B), but for the whole chromosome [dashed lines from (B) to (C)]. (D) Same whole chromosome as in (C), but with genetic (not physical) distance along the x axis. (E and F) Genomewide SNP divergence between marine and freshwater ecotypes globally and in the northeastern Pacific basin, with specific-threshold EcoPeaks in red. (G) Many differentiated regions overlap the location of major quantitative trait loci (QTLs) controlling various morphological, physiological, and behavioral traits in previous genetic crosses [percent variance explained (PVE) > 20, interval < 5 Mb from Peichel and Marques (53)].
genomic feature enabling extremely fast evolution of freshwater phenotypes in stickleback from the northeast Pacific basin.

The *Eda* SNP associated with armor plate variability (chrIV: 12,823,875 T>G (49)) is within the second most significant specific TempoPeak on chrIV. In both CH and SC, the G allele increases rapidly from an initial frequency of <1% to over 50% within 8 years, while approaching fixation in LB by 15 years. Notably, the square of G-allele frequencies (i.e., the expected number of GG homozygotes) tracks closely with frequencies of the low–armor plate phenotype, consistent with almost complete recessiveness (*h* = 0.0) for the G allele for this phenotype (Fig. 2B). Nonetheless, to fit the allele frequency trajectory of this SNP, and, in particular, the extremely rapid
increase in CH and SC, it was necessary to impose a dominance coefficient \((h)\) of 1.0 along with a very large selection coefficient \((s)\) of 0.55, as in a recent paper focusing on this locus \((18)\).

Like Eda, most TempoPeaks display similarly sharp left-shifted sigmoidal allele frequency trajectories, indicating very strong and dominant-positive selection (Fig. 2C and section S20). When modeling each peak SNP as independent, we find an extremely high mean \(s\) of 0.30 (5th, 95th percentile 0.08 to 0.53) and \(h\) of 0.98 (5th, 95th percentile 0.95 to 1.0) for the 344 specific TempoPeaks found in CH + SC + LB. The estimated \(s\) values for chrIV, where there are 69 TempoPeaks, are particularly high (mean \(s = 0.38\), consistent with the accelerated evolution of this whole chromosome observed via a chromosome-wide \(F_{ST}\) analysis comparing the founding generation of CH, SC, and LB to all subsequent years (section S15).

**Features associated with EcoPeak evolution**

The remarkable speed at which northeast Pacific stickleback adapt to new freshwater environments suggests that analysis of EcoPeaks may provide unique insights into optimal genomic properties for evolution. Using *Gasterosteus nipponicus*, *Gasterosteus wheatlandi*, and *Pungitius pungitius* for calibration, we estimated molecular divergence time between a pair of freshwater (Little Campbell upstream) and marine (Little Campbell downstream) stickleback in windows tiled across the genome (Fig. 3A). Although peaks shared globally trend older than those found just within the northeast Pacific (1800 ka versus 1600 ka, \(P = 0.18\), the imputed ages overlap considerably (Fig. 3A). We estimate that the majority (161 of 209) are over a million years old and have cycled between freshwater and marine environments many times during this long history, likely persisting at high frequency in freshwater habitats south of the zone of glaciation during the Ice Ages and at more northerly latitudes during previous interglacials and the Holocene.

Contrary to our expectations that recombination would disassemble regions over time, we found that older EcoPeaks are larger than younger ones (Fig. 3B). This signature is strongest at the most significant markers within each EcoPeak, which are typically older than more distal sequences (Fig. 3C). This suggests that individual regions may grow over time, with alleles originally based on an initial
beneficial mutation accumulating additional linked favorable mutations, snowballing over time to form a finely tuned haplotype with multiple adaptive changes. This is consistent with work in other species identifying examples of evolution through multiple linked mutations that together modify function of a gene (50–52) and implies that progressive allelic improvement may be common.

We also observed that EcoPeaks frequently overlap major quantitative trait loci (QTLs) in stickleback [73 of 209 overlaps observed versus 32 of 209 expected, \( P < 1 \times 10^{-15} \); Fig. 1G (53)], suggesting that these variants underlie many mapped phenotypic traits. Just as the QTLs cluster in “supergene” complexes (54), so too do EcoPeaks (median observed interpeak distance 192 kb versus 795 kb expected, \( P = 4.88 \times 10^{-10} \)). One particularly large complex (chrIV: 8 to 17 Mb) contains 22 EcoPeaks and the major QTLs controlling many aspects of both defensive armor and trophic morphology (e.g., the length of dorsal and pelvic spines, the number of armor plates through Eda, gill rakers, and teeth). Thus, clustering may have important functional effects by allowing multiple traits and underlying EcoPeaks to be selected and inherited as a single unit, especially when in tight linkage. A fine-scale recombination map of RS stickleback (generated with LDhelmet (55)) shows that EcoPeaks are highly enriched in regions of low average recombination, forming tightly linked haploblocks (Fig. 3D, compare Fig. 1, C and D; section S14). EcoPeaks are also enriched near local recombination hotspots within their neighborhood (Fig. 3E), potentially facilitating reassembly of larger haplotype blocks upon freshwater colonization (also see section S19).

To further examine the frequency and size of haploblocks in individual fish, we surveyed 1643 stickleback from three Alaskan marine populations by SNP array genotyping (sections S5 and S12). While most marine fish heterozygous for freshwater alleles carry a relatively small haploblock, some carry multi-megabase haploblocks containing multiple EcoPeaks (Fig. 3F). Thus, a proper treatment of rapid stickleback evolution needs to account for the complex linkage of EcoPeaks rather than treating them independently.

**Modeling the genomic landscape of contemporary evolution**

To estimate a more realistic distribution of fitness effects (DFE) that incorporates the genome’s recombination landscape, we developed a deep neural network (DNN) approach that uses forward simulations (section S21). Our simulations, which are conceptually similar to those of Galloway et al. (56), attempted to replicate the dynamics of the “transporter model” (29), with one large \( (N_e = 10,000) \) anadromous population connected independently by gene flow to 10 smaller \( (N_e = 1000) \) established freshwater populations. After 1000 generations, we founded three new freshwater populations from the anadromous population, thus generating simulated allele frequency trajectories that reflect our annual LB, CH, and SC samples (Fig. 4A).

Focusing our DNN analysis on a subset of 19 specific TempoPeak SNPs separated by \( \geq 0.4 \) cM (\( \approx 100 \) kb) along chrIV, we closely replicated observed allele trajectories of positively selected freshwater alleles across all SNPs simultaneously using a beta distribution–shaped DFE, for which the mean \( s \) across the 19 TempoPeaks was 0.063 and the standard deviation was 0.030, with reciprocal fitness costs implemented in the marine population (Fig. 4C). The estimated \( s \) from our DNN was thus substantially smaller than the mean of 0.48 when each SNP was considered independently. In addition, 18 of 19 SNPs were predicted to be fully dominant and none fully recessive under the best model.

We validated our best-fit DNN model by simulating the 19 selected TempoPeaks SNPs with the estimated DFE along with \( \sim 400 \) neutral SNPs distributed randomly along chrIV. Despite the neutral SNPs not being used in training the DNN, we were able to mimic the overall topology of the CMH scores across the entire genome, suggesting that our model was capturing the overall genomic architecture of freshwater adaptation (Fig. 4D). Our best-fit DNN model also appeared to recapitulate much of the haplotype structure of the array data from individuals from RS, LB1999, and LB2013 (Fig. 4B). Notably, the transition to freshwater alleles appears to be somewhat slower on the right half of chrIV, where there are fewer EcoPeaks, TempoPeaks, and QTLs, and this difference was observable in both the empirical and simulated data.

Overall, our model suggests that extremely rapid and replicable allele frequency increases on chrIV in LB, CH, and SC are mostly driven by multiple linked (primarily) dominant alleles, each with relatively smaller \( s \) values that act in concert, with recombination hotspots between them (section S19) allowing rapid reassembly of optimum freshwater haplotypes, consistent with the transporter hypothesis. The lower individual \( s \) values may allow these dominant alleles to persist in the marine environment at low frequency after being disassembled by recombination, especially if some act in epistasis.

**Biological features with predictive power**

Given the genomewide dynamism of the earliest stages of freshwater adaptation, we attempted to identify genomic features that predict the speed of evolution at TempoPeaks and understand why some peaks are consistently selected more rapidly than others (section S22). We used CMH scores as a proxy of evolutionary speed for each TempoPeak in CH + SC + LB and regressed these against a variety of sequence features.

The best predictor for the speed of evolution is the degree of ecotypic differentiation between marine and long-established freshwater populations (Pacific EcoPeak \( P \) value), with variants more commonly differentiated in the northeast Pacific being selected more quickly (Fig. 5A and fig. S81). Fisher’s geometric model indicates that alleles with large effects are usually disfavored; however, the “prefiltering” of ancient SGV that counters this tendency (12) largely benefits alleles that are broadly positively selected, possibly explaining this result.

We also found that larger TempoPeaks are typically selected more rapidly. Similarly, greater TempoPeak density predicts more rapid divergence, suggesting that our simulation accurately reflects how nearby loci mutually reinforce their collective selection. Overlap with major QTLs also has a strong association with rapid evolution, while other variables such as increased sequence divergence, decreased recombination rate, increased gene overlap, increased sequence conservation, increased Ka/Ks, and decreased ancestral marine frequency have smaller contributions to predictive power for speed of selection (Fig. 5A).

We also tested whether underlying sequence characteristics could predict not only the speed of selection in CH + SC + LB but also the location of the selected regions themselves (section S23). Recombination rate, QTL overlap, allelic age, and an integrated genomic context score (section S23) that incorporate the previous features are all useful predictors (Fig. 5B). By combining these fundamental features into a logistic model trained on the survey of extant populations, the most confident predictions of selected regions in the rest of the genome achieve 85% precision. This model performs 67% as well as predictions based only on empirical repeatability in extant populations in the northeast Pacific (Fig. 5B). Thus, our understanding of underlying principles reflects an incomplete yet substantial proportion of evolutionary repeatability.
Fig. 4. DNN simulation–based modeling of rapid and repeated stickleback evolution. (A) Schematic showing evolutionary model of forward simulations under the transporter hypothesis. Red horizontal bars, anadromous (AN) ancestor; blue circles, descendant freshwater isolates; red to blue shaded circles, three adapting freshwater populations (i.e., LB, CH, and SC) founded recently by anadromous stickleback; and arrows, gene flow or founding events. (B) Genotypes across chrIV for freshwater-associated SNPs in RS (n = 750), LB in 1999 (n = 25), and LB in 2013 for (left) observed and (right) simulated data under best-fit DNN model. Anadromous homozygous, red; heterozygous, yellow; and freshwater homozygous genotypes, blue, respectively. (C) Allele frequency trajectories for LB, CH, and SC in 100 simulations under the best-fit DNN model for five randomly selected SNPs. Larger points, observed data. (D) Distribution of average CMH scores in windows of 2500 bp across chrIV for (top) observed and (bottom) simulated data under best-fit DNN model. Red dotted lines, locations of SNPs under selection and used to fit DNN.
Parallels in distant species
To test the generality of these predictive factors, we applied the stickleback-trained model to a dataset of 12 pairs of species of Darwin’s finches (section S23) (57). Darwin’s finches have undergone adaptive radiation in the Galápagos Islands over the last several hundred thousand years, are ~435 million years divergent from stickleback, and face very different selective pressures. As in stickleback, however, the “islands of divergence” of all 12 analyzed pairs of species of Darwin’s finches (sensu Han et al.) are enriched for ancient alleles overlapping mapped QTLs with low recombination rates. The top 100 windows predicted by the stickleback model recover a median of 28-fold more previously identified islands of divergence than expected by chance ($P < 1 \times 10^{-10}$; Fig. 5C), including the $Alx1$ and $Hmga2$ loci implicated in beak morphology in multiple species pairs (even without QTL input). The model also recovers a substantial proportion of differentiated loci in a recent case of cichlid speciation (58). Thus, a handful of basic genomic properties allow strong quantitative predictions of the location of key evolutionary loci, even across widely separated branches of life.

DISCUSSION
The importance of SGV for evolution is becoming increasingly apparent, especially in species with large genome sizes (59), including humans (60). At first glance, the dependence of threespine stickleback on SGV for freshwater adaptation may appear to be a peculiarity in terms of repeatability and speed and their particular natural history. However, by more comprehensively understanding the dynamics of this highly optimized process, we have extracted general features of genome architecture and evolution that successfully translate to species on distant branches of the tree of life, thus demonstrating the tremendous power of the stickleback system to identify unifying principles that underlie evolutionary change.

MATERIALS AND METHODS
Sample collection and DNA preparation
Fish for all downstream genomic analyses were trapped following Institutional Animal Care and Use Committee (IACUC) guidelines using unbaied minnow traps set near shore, immediately euthanized with MS-222 (tricaine methanesulfonate), and then preserved in 70 or 95% ethanol (section S1). DNA extraction was performed using either phenol:chloroform isolation (61) and quantified using a NanoDrop spectrophotometer or using the DNeasy 96 Blood & Tissue Kit following the standard “animal tissue” protocol and quantified using the Qubit High-sensitivity DNA Assay (section S2). Equimolar pooling of samples from RS, CH, SC, and LB was performed using an Opentrons OT-2 robot (section S3).
Genome sequencing and genotyping

Samples from long-established populations underwent WGS sequencing on a HiSeq 2000 using 2 × 76-bp paired-end sequencing libraries. Contemporary pools were sequenced on either a NovaSeq 6000 or Illumina HiSeq 2500 using 2 × 150-bp paired-end sequencing libraries. Contemporary WGS was performed by Beijing Genomics Institute using their proprietary DNBseq technology using 2 × 100-bp paired-end libraries (section S4). A custom Illumina 384 GoldenGate array was designed for SNP genotyping (section S5).

Bioinformatic processing

We constructed a slightly modified reference genome based on the recent Hi-C–guided improvement of the stickleback genome (62), to which we refer as gasAcu1-4, that includes a new chrP and a new mitochondrial genome (section S6). Reads were mapped to gasAcu1-4 using bwa mem (63) and Picard was used to add read groups and mark duplicate reads. Indel realignment and base quality recalibration were performed using Genome Analysis Toolkit (GATK) (64). HaploTypeCaller and GenotypeGVCFs were used for variant calling for WGS data. Allele frequencies in pool-seq populations were calculated using the maximum-likelihood method of Lynch et al. (65) and PoPoolation2 (v1201) (section S7) (66).

Analysis

EcoPeaks were identified using two approaches, the first following the same genetic distance–based approach as Jones et al. (8) based on 2500-bp windows sliding every 500 bp, and the second analyzing the distribution of allele counts between marine and freshwater populations at every base position in the genome with two alleles present at >10% frequency in the combined analysis metapopulation. For both the SNP-based and 2500-bp window–based analyses, nearby significant values were grouped into the EcoPeaks that behaved as a single unit using a greedy algorithm. Peaks were filtered at either a 1% FDR for the specific calls or at 5% for the sensitive calls. The single base and window peaks were then intersected for the final specific calls or unioned for the final sensitive calls (section S9). Allelic divergence and age were computed from five upstream (freshwater) and five downstream (marine) fish from Little Campbell River. Nonoverlapping 1-kb windows were tiled across the genome, and variants homozygous for different alleles were counted and used to compute marine-freshwater sequence divergence $d$ (section S11). A rho-based recombination map was constructed on the basis of 20 RS genomes using LDhelmet (55) following a similar methodology to that of Shanfelter et al. (67), though with some minor modifications, and converted to genetic distance based on the pedigree-based linkage map generated by Glazer et al. (68) (section S14). $F_{ST}$ was calculated for each pool-seq population against its youngest counterpart using the ratio of averages method implemented by Bhatia et al. (69) (section S18). We used a modified CMH test (48) to identify SNPs that had shown a significant change in allele frequency in our contemporary time-series data (section S17). We followed the same general EcoPeak identification methodology to define TempoPeaks from our CMH $P$ values. Sensitive TempoPeaks were based on a 5% Bonferroni-corrected $P$ value threshold merging SNP and window-defined peaks. Specific TempoPeaks were based on a 1% $P$ value threshold only considering window-defined peaks (section S18). We applied the deterministic method described by Taus et al. (70) to estimate $s$ for the SNP with the largest CMH score in each significant TempoPeak. We estimated $s$ and $p0$ (initial allele frequency), both assuming that the dominance coefficient $h$ is 0.5, as well as simultaneously estimating $s$, $p0$, and $h$ (section S20). We additionally developed a DNN approach to estimate the DFE of multiple linked TempoPeaks on chrIV. This analysis includes three main stages: (i) simulating linked loci with positive selection parameterized with various randomly drawn DFEs within the context of the demographic and evolutionary model of the transporter hypothesis, (ii) using the DNN framework to estimate the best DFE given the observed allele frequency trajectory data from our pool-seq experiments, and (iii) comparing the transporter hypothesis under the best-fit DFE to various features of the observed genomic data (section S21). Genomic features of interest were then used to predict speed of TempoPeak selection in the contemporary evolution populations using a linear regression model (section S22). We also applied genomic features of interest to predict the genomic loci of selection in the contemporary evolution populations via multivariate logistic regression. Last, we applied this model with the same terms and weights to datasets from Darwin’s finches (57) and Lake Victoria cichlids (58) using the previously published analyses as our truth sets (section S23).

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at http://advances.sciencemag.org/cgi/content/full/7/25/eabg5285/DC1

REFERENCES AND NOTES

1. C. Darwin, Fertilisation of Orchids (Murray, 1862).


landscapes between a
improvement and
The
S. threespine stickleback.

reduction in
Y. B.
Duplication, deletion, and
T. D590–D598
Nucleic Acids Res.
B.
W.
identification, and
M.
S.
D.
A.
in
contributes to
S.
C.-J.
observed in
R.
and
P.

A.

P. Purcell, B.
Arif, W.
Balabhadra, D.
M. J.
Nag, M.
García Girón, T.
F.
Schluter, D.
J.
R.
C.
Osoegawa, K.

D.
A.

Purcell, C. Chang, PLINK: www.cog-genomics.org/plink/1.9/.

N. Patterson, A. L. Price, Population structure and


Funding: This was supported by NSF BSR9005758, BSR9004619, DEB0211391, DEB0322818, DEB0509070, and DEB0919184 to M.A.B.; NIH R01GM124330 to K.R.V.; DFG SPP1819 and ERC 617219 to F.C.J.; by Newcomb Institute grants to D.C.H.; NSF GRFP to G.A.R.K.; NSF GRFP 1656518 and NIH ST22GM007790 and Stanford CEHG Fellowship to H.I.C.; and NIH 5P50HG002568 and 3P50HG002568-09S1 to D.M.K. D.M.K. is an investigator of HHMI.

Author contributions: M.A.B: experimental design, Alaskan sampling and population founding, tissue sampling, and morphological data collection. P.J.P.: Alaskan sampling and population founding. F.A.V.: Alaskan sampling and logistical support. W.E.A.: Alaskan sampling and morphological data collection. D.C.H.: Alaskan sampling. G.A.R.K.: analysis of geographic populations and genomic properties predictive of evolution. D.N.Y., M.M., and T.S.B.: DNA extraction, quantitation, selection, and pooling of Alaskan samples. F.C.J.: design of geographic sampling and SNP array and SNP array analysis. S.D.: sample curation, DNA preparation, and SNP array calling. D.M.A. and R.M.M.: SNP array genotyping. M.K.: SNP array analysis. H.I.C.: transfer and visualization of genome annotations. F.D.P.: geographic population sequencing. D.M.K.: experimental design and conceptual guidance. K.R.V. and K.R.: analysis and modeling of contemporary populations. G.A.R.K. and K.R.V. wrote the manuscript with input from all authors. Competing interests: The authors declare that they have no competing interests. Data and materials availability: All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. All WGS Illumina data of extant populations have been deposited in the Sequence Read Archive (www.ncbi.nlm.nih.gov/sra) at accession PRJNA247503. All WGS Illumina data for contemporary pool-seq experiments have been deposited in the Sequence Read Archive at accession PRJNA671824. All WGS BGI data for RS 2009 genomes used to construct the recombination map have been deposited in the Sequence Read Archive at accession PRJNA671690. SNP genotyping array data and ancestral reference sequence have been deposited on Data Dryad (https://doi.org/10.5061/dryad.pvmcvdrgm), EcoPeak, TempoPeak, and RS recombination rate data can be visualized and downloaded (via the Table Browser (71) of the UCSC Genome Browser (http://genome.ucsc.edu/) (72)) by copying the following track hub (73) URL into the “My Hubs” tab at https://genome.ucsc.edu/cgi-bin/hgHubConnect?https://sbwdev.stanford.edu/kingsleyAssemblyHub/hub.txt. The assembly hub must be opened through the UCSC Genome Browser, not directly. We provide the gasAcu1-f reference genome as well as liftOver chains for converting to and from the original Broad 51 stickleback reference genome (gasAcu1) via Data Dryad at https://doi.org/10.5061/dryad.547d7wm6l.

Submitted 11 January 2021
Accepted 5 May 2021
Published 18 June 2021
10.1126/sciadv.abg5285

Predicting future from past: The genomic basis of recurrent and rapid stickleback evolution


Sci Adv 7 (25), eabg5285.
DOI: 10.1126/sciadv.abg5285