nature ecology & evolution

ARTICLES

https://doi.org/10.1038/s41559-020-1201-y

Check for updates

Evolutionary origins of genomic adaptations in an invasive copepod

David Ben Stern and Carol Eunmi Lee ✉

The ability of populations to expand their geographical ranges, whether as invaders, agricultural strains or climate migrants, is currently one of the most serious global problems. However, fundamental mechanisms remain poorly understood regarding factors that enable certain populations, such as biological invaders, to rapidly transition to novel habitats. According to one hypothesis, environmental fluctuations in the native range could promote successful invasions by imposing balancing selection on key traits and maintaining the genetic variation that enables rapid adaptation in novel habitats. Here we test the genomic predictions of this hypothesis by performing whole-genome sequencing of multiple independent invasive freshwater and native saline populations of the copepod *Eurytemora affinis* complex. We found that invasive populations have repeatedly responded to selection through the parallel use of the same single-nucleotide polymorphisms and genomic loci, to a much greater degree than expected. These same loci were enriched for signatures of long-term balancing selection in the native ranges, with 15–47% of loci exhibiting significant signatures of balancing selection. The strong association between parallel evolution in the invaded range and balancing selection in the native range supports the hypothesis that fluctuating habitats can promote invasive success and that balancing selection might serve as a widespread and important mechanism that enables rapid adaptation in nature.

Invasive species pose one of the greatest threats to biodiversity, ecosystem integrity, agriculture, fisheries and public health, with economic costs amounting to hundreds of billions of dollars per year worldwide[1,2]. Global climate change is projected to increase the number and impact of invaders in an unprecedented and complex manner[3–8], requiring a comprehensive understanding of the mechanisms that facilitate successful biological invasions[9–12]. A longstanding debate has focused on the precise factors that generate successful invaders, given that an exceedingly small proportion of introduced species are able to establish in new habitats and then become invasive[13]. Numerous hypotheses have been proposed and tested, including the roles of propagule pressure, transport opportunity, habitat matching, fecundity and population size. However, these hypotheses have not found consistent empirical support across taxonomic groups and invasion events, offering limited powers of prediction[14–18].

Lee and Gelembiuk[19] proposed an evolutionary mechanism that could promote the emergence of invasive populations and hypothesized that the selection regime in the native range acts as a crucial factor that affects invasive success[19]. They observed that invasive populations tend to originate from habitats marked by disturbance or temporally varying conditions[19,20]. Consequently, they hypothesized that many invasive populations have originated from native populations undergoing balancing selection, resulting from fluctuating environmental conditions. This mechanism would tend to operate in organisms with short generation times, relative to the period of environmental fluctuations, such that different alleles would be favoured by selection in different generations[19]. Such a selection regime could maintain standing genetic variation in the native range and provide the genetic substrate upon which positive selection could act during invasions[10,15,17,21–24]. However, this hypothesis had not previously been tested empirically.

Balancing selection is a form of natural selection that favours more than one allele at a locus, and its ability to maintain standing genetic variation has remained a contested topic in evolutionary biology[24–30]. In particular, the conditions under which temporally fluctuating selection can maintain polymorphisms through time have been thought to be fairly restricted[31–35]. Moreover, little empirical evidence exists regarding the extent to which adaptation to novel habitats could be facilitated by balancing selection in the native range[24,36–38]. In theory, balanced genetic variants might have a higher probability than neutral variation of contributing to adaptation to novel habitats, as balancing selection can maintain variants at relatively high frequencies and increase their fixation probabilities under new selection pressures[24,39–42]. Recently, several studies of adaptation from standing genetic variation have found signatures of directional selection acting on alleles that segregate at intermediate frequencies, suggesting the presence of some form of balancing selection[38,42–45]. However, whether alleles under directional selection in colonizing or invasive populations could arise from those maintained under balancing selection in their native ranges remains largely untested.

Moreover, balancing selection in the native range could increase the chances of selection acting on the same loci during replicated invasion events. When closely related populations are independently exposed to the same novel selection pressure, adaptation could potentially proceed from shared standing variation[46]. The prevalence of parallel selection acting on shared standing variation depends on the divergence time between populations and the factors that influence the retention of ancestral variation[47]. Fluctuating selection due to shared environmental forces in the native range could result in multiple native range populations harbouring the same ancestral genetic variation. The presence of these shared balanced variants would increase the probability of genetic parallelism during invasions, thereby enhancing the predictability of evolutionary responses to environmental change[48].

The common estuarine and saltmarsh copepod *E. affinis* complex provides an excellent model system in which to explore these

Department of Integrative Biology, University of Wisconsin-Madison, Madison, WI, USA. ✉e-mail: carollee@wisc.edu

Content courtesy of Springer Nature, terms of use apply. Rights reserved

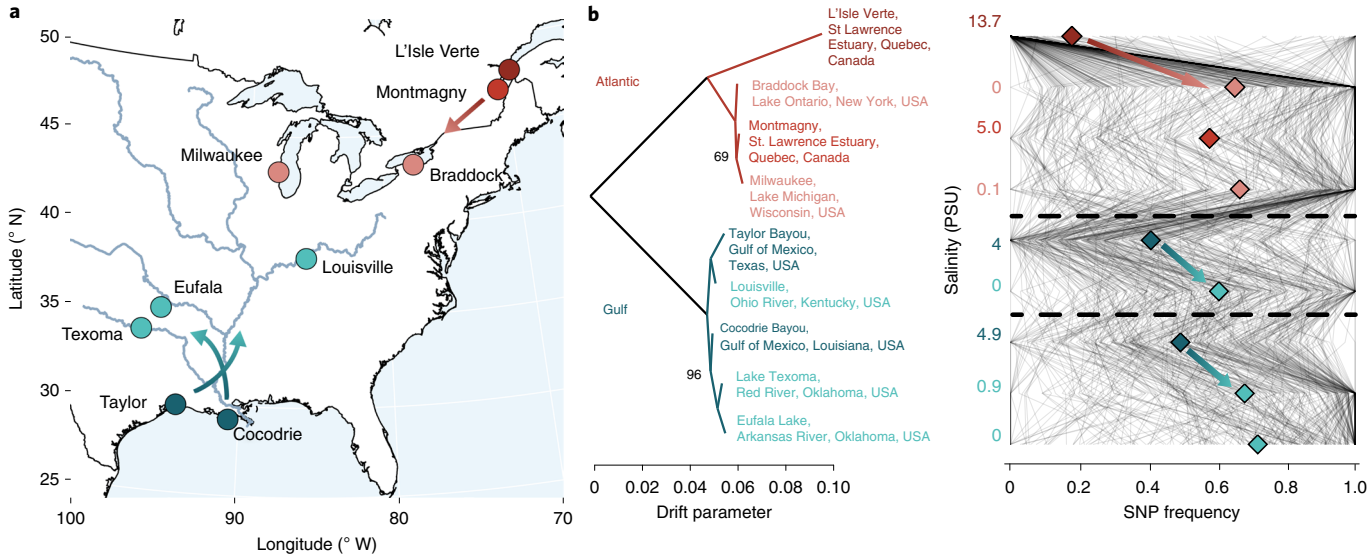


Fig. 1 | Population genomic signatures of parallel freshwater invasions. a, Map of sampling locations of *E. affinis* complex populations in North America. Invasive freshwater populations are represented by light-coloured circles and native saline populations are represented by dark-coloured circles. Red circles represent populations from the Atlantic clade and green circles represent populations from the Gulf clade[54]. Arrows indicate the inferred directions of independent freshwater invasions. **b**, Parallel frequency shifts associated with repeated saline to freshwater invasions (parallel candidate SNPs, Supplementary Box 1). The population phylogeny was estimated from SNP frequency correlations using TreeMix v.1.13. Invasive freshwater populations are shown as light colours and native saline populations are shown as dark colours. All nodes have bootstrap support of 100%, except those that are shown. SNP frequencies (grey lines) are shown for parallel candidate SNPs ($n=347$) with both significant signatures of directional selection (BayeScan 3) and association with salinity (BayPass). SNP frequencies are polarized to keep directionality consistent among SNPs. Diamonds represent the mean candidate SNP frequencies of the corresponding population on the phylogeny. The numbers next to population names indicate the salinity at which each population was sampled. Dotted horizontal lines delineate clades with both invasive and native populations, indicating the presence of recent independent invasions. The arrows represent the direction of SNP frequency shifts between saline and freshwater populations.

questions. Within the past approximately 70 years, populations from this species complex have been very successful at invading freshwater habitats, mediated by human activity. Due to increases in shipping and ballast water discharge in recent years, *E. affinis* complex populations have invaded freshwater habitats multiple times independently, from genetically divergent clades throughout the Northern Hemisphere[20,49] (Fig. 1). These replicated invasions enable the discovery of genomic signatures that are consistently associated with invasion success. Notably, native range saline populations of the *E. affinis* complex exhibit several properties that could greatly expand the conditions under which balancing selection can maintain polymorphisms, including seasonal fluctuations in salinity, overlapping generations in the form of diapause egg banks[50–52] and beneficial reversal of dominance with respect to salinity tolerance[53] (see 'Discussion' and Box 1). Thus, balancing selection might provide a plausible mechanism that facilitates invasions in this system, and potentially many other systems that give rise to invasive populations[19].

Thus, in this study, we took advantage of the independent invasions and rapid parallel physiological evolution in the *E. affinis* complex to test the hypothesis that fluctuating habitats promote invasion success by imposing balancing selection on key traits and associated genomic loci. To test this hypothesis, we (1) explored genomic targets of directional selection during independent saline to freshwater invasions and assessed the prevalence of molecular parallelism across the genome, and (2) examined whether the specific genomic targets of selection in the invasive populations exhibit signatures of balancing selection in the native populations.

To address our hypotheses, we performed whole-genome sequencing of 100 pooled individuals per population for multiple invasive populations and their respective native range populations (Fig. 1a). This evolutionary replication gave us greater power to determine the specific loci that undergo both directional and balancing selection. We illustrate that the processes and mechanisms that underlie freshwater invasions can be highly predictable, both in terms of generating successful invaders in the native range and inducing parallel evolutionary responses in the invaded range. If indeed the same loci that undergo parallel selection in the invaded range are also under balancing selection in the native range, such information could offer unprecedented powers of prediction on which populations have the capacity to invade.

Results

Population structure and history. Our population genomic data support the occurrence of multiple independent invasions from genetically divergent source populations, corroborating previous studies using mtDNA[49,54] (Fig. 1a). The deep phylogenetic split between the Atlantic and Gulf clades was evident from the small proportion of shared variation present and the structure of the population phylogeny built from the variance–covariance matrix of single-nucleotide polymorphism (SNP) frequencies[55] (Fig. 1a). In terms of shared variation in the two clades, only 5.52% of the global 6,635,765 biallelic SNPs had a minor allele frequency (MAF) > 0.05 in both clades. In addition, we found much greater SNP differentiation among populations in different clades (mean $F_{CT}=0.19$) than within clades (mean $F_{SC}=0.04$) (estimated using BayeScan 3 (also known as BayeScanHierarchical)[56], see 'Widespread genomic signatures of parallel directional selection during freshwater invasions'). The population phylogeny indicated at least two independent freshwater invasions in the Gulf clade and one invasion in the Atlantic clade. Genome-wide genetic diversity estimates (θ) (Supplementary Table 2) were not significantly lower in the invasive populations (phylogenetic generalized least squares; $\theta_{Watterson}$ ~ salinity, $t=-0.167$, d.f. = 9, $P=0.872$; θ_{π}~salinity: $t=-0.335$, d.f. = 9, $P=0.748$; see Methods), indicating the lack of population bottlenecks following invasion events.

Content courtesy of Springer Nature, terms of use apply. Rights reserved

Box 1 | Factors that could promote balancing selection in native populations of the *E. affinis* complex

Fluctuating environmental conditions Native populations of the *E. affinis* complex experience seasonal fluctuations in salinity ranging from 5 to 40 PSU[20]. With around six generations per year, the period of salinity fluctuations is greater than generation time, exposing different generations to different selection pressures throughout the course of the year.

Negative genetic correlations In the *E. affinis* complex, negative genetic correlations (consistent with antagonistic pleiotropy) exist between saltwater and freshwater tolerance[99–101]. Thus, under seasonally fluctuating salinities, saline and freshwater tolerance would be favoured by selection at different times and different generations (with around six generations per year). Such fluctuating selection could promote the maintenance of genetic variation, provided that this polymorphism can remain protected against strong negative selection[102–104].

Beneficial reversal of dominance Beneficial reversal of dominance (BRD) with respect to salinity tolerance has been demonstrated in the *E. affinis* complex[53]. BRD is the phenomenon in which alternate alleles are always dominant in the environment in which they have a higher fitness. Under fluctuating selection, BRD can greatly increase the fitness of heterozygotes and protect polymorphisms, as alleles will have reduced exposure to selection during conditions in which they are not beneficial[70,102–107]. Currently, BRD has been demonstrated in only the *E. affinis* complex[53], although there is evidence that dominance in gene expression may be environmentally dependent in *Drosophila*[78].

Overlapping generations Populations of the *E. affinis* complex have overlapping generations generated through diapause egg banks[50–52], which can protect polymorphisms against fluctuating selection over time[102,105].

Phenotypic plasticity Populations of the *E. affinis* complex exhibit plasticity in physiological tolerance and performance associated with salinity changes in the native range[101]. Plasticity has been predicted to contribute to the ability of balancing selection to maintain polymorphisms through a genomic storage effect[108,109].

Widespread genomic signatures of parallel directional selection during freshwater invasions. Genomic signatures of repeated freshwater invasions were enriched for signatures of directional selection both on the same SNPs and on different SNPs in the same genomic windows in both clades (see Supplementary Box 1 for definitions of relevant terms). This pattern of selection on many of the same loci was surprising given the deep phylogenetic split that we estimated between the two distinct clades of the *E. affinis* complex (Fig. 1a). A substantial number of SNPs and genomic windows displayed signatures of selection in only one clade, suggesting some independent genomic routes to freshwater adaptation. However, using both neutral simulations and window randomizations to generate null distributions, we found that the genome-wide signatures of repeated evolution at the same loci were significantly greater than expectation (Supplementary Information section I).

To detect SNPs associated with freshwater invasions, we performed genome-wide scans for directional selection and association with salinity with BayeScan[56,57] and BayPass[58], respectively. A significant 'association with salinity' indicates that the shift in allele (SNP) frequency was correlated with changes in salinity, suggesting a functional relationship between the rise of particular alleles and freshwater adaptation. In addition to detecting genomic signatures of freshwater invasions in each clade separately, we sought to detect signatures of directional selection (Supplementary Box 1) found in common in both clades with the goal of increasing power to detect loci truly associated with freshwater adaptation[46,59,60]. We used two approaches to detect signatures of selection found in common between the two clades given their apparent divergence: (1) we tested for signatures of parallel directional selection and association with salinity on shared SNPs (to uncover parallel candidate SNPs; Supplementary Box 1) and (2) detected 10-kb genomic windows that overlapped between the two separate genome scans for selection in each clade (shared candidate windows; Supplementary Box 1). The first analysis was designed to detect signatures of parallel frequency shifts at exactly the same SNPs, whereas the second analysis was designed to detect small genomic regions that contained shared targets of selection. Detecting shared candidate windows could capture different SNPs under selection that occur at the same loci in different lineages. As a result of our analyses, we obtained four sets of candidate loci (Supplementary Box 1).

To assess the extent of parallel frequency shifts in shared SNPs, we used a hierarchical *F*-model (BayeScan 3)[56] to compare statistical support for three selection models and a neutral model for all SNPs with a MAF > 0.05 in both clades ($n = 366{,}781$). The three selection models were constructed to determine whether a SNP had a signature of directional selection in only the Atlantic clade, in only the Gulf clade, or in both clades in parallel. A substantial proportion of SNPs with significant signatures of directional selection showed the highest support for the parallel selection model, rather than selection in either clade alone (parallel, 42.5%; only the Atlantic clade, 19.5%; only the Gulf clade, 38.0%). In terms of the numbers of significant SNPs, 2,970 SNPs displayed signatures of parallel directional selection out of a total of 6,981 SNPs with signatures of directional selection in at least one clade. Of the 2,970 SNPs with signatures of parallel directional selection, 349 SNPs also showed significant association with salinity across all populations. We removed 2 of these 349 parallel candidate SNPs (Supplementary Box 1) from downstream analyses because they showed evidence of potential copy-number variation (Supplementary Information section II). We found that genome-wide signatures of parallelism for shared SNPs were significantly greater than the degree of parallelism expected under genetic drift alone, based on neutral simulations (Supplementary Information section I, Supplementary Table 3). Specifically, our dataset contained 29 times the number of SNPs with signatures of parallel selection and association with salinity than expected under drift alone (Supplementary Information section I).

The analyses of directional selection and association with salinity in each separate clade yielded 679 and 2,092 non-parallel candidate SNPs in the Atlantic and Gulf clades, respectively. The greater number of non-parallel candidate SNPs found in the Gulf clade was likely due to the power gained from the additional population and invasion event sampled in the Gulf clade relative to the Atlantic clade (Fig. 1). Notably, the number of shared candidate windows ($n = 279$; Supplementary Box 1) around significant SNPs was significantly greater than expected using window randomization ($n = 120.31 \pm 0.204$, $P < 0.0001$; Supplementary Information section I). Thus, the genome-wide signature of selection on the same small genomic regions was 2.3-fold greater than expected by chance given the size of the windows and genome. While the greater number of non-parallel, compared with the number of parallel, candidate SNPs could indicate largely different genomic routes to freshwater adaptation, the greater than expected signatures of selection both at the same SNPs and at different SNPs in the same 10-kb windows suggest that selection has often acted on the same loci across repeated invasion events and that these parallel loci represent important genomic signatures that underlie the adaptation to freshwater habitats.

Content courtesy of Springer Nature, terms of use apply. Rights reserved

Table 1 | SNPs within ion-transporter genes showing signatures of directional selection associated with freshwater invasions

Candidate loci	Gene names of the ion transporters
Parallel candidate SNPs	*NHA*, paralogues 3, 4, 5, 7; *NKA*, subunit α, paralogue 2; *NKA*, subunit β, paralogue 5; *Rh*, paralogue 2
Shared candidate windows	*NHE*, clade X, paralogue c; *NHA*, paralogue 6; *NKA*, subunit α, paralogue 2
Non-parallel candidate SNPs, Atlantic clade	*NHA*, paralogues 1, 2, 4, 6, 7; *NKA*, subunit α, paralogues 2, 4, 5; *NKA*, subunit β, paralogue 5; carbonic anhydrase (*CA*), paralogues 1, 5, 12
Non-parallel candidate SNPs, Gulf clade	Ammonia transporter (*AMT*), paralogue 3; *Rh*, paralogue 2; *NKA*, subunit α, paralogues 2, 6; Na^+,K^+,2Cl^- cotransporter (*NKCC*), paralogues 3, 4

See Box 1 for definitions of the terms.

Ion transporter genes are overrepresented as targets of selection. Ion transport and related Gene Ontology (GO) terms were overrepresented in our candidate SNPs (Supplementary Table 4). Among significant GO terms (false-discovery rate (FDR)-adjusted $P < 0.05$), 37.5% (3 out of 8) and 55.5% (5 out of 9) were related to ion transport in the set of parallel candidate SNPs and non-parallel candidate SNPs of the Atlantic clade, respectively (Supplementary Table 4). Top GO categories in the parallel set included several ion transporter terms, such as 'sodium ion transport', 'sodium-proton antiporter activity', 'lithium-proton antiporter activity' and 'regulation of intracellular pH'. Ion transport had previously been implicated in freshwater adaptation in the *E. affinis* complex[61–63], and here our GO analysis also implicates ion transport as the dominant physiological function associated with freshwater invasions. Other top GO terms in the set of parallel candidate SNPs included several terms related to gene regulation (for example, 'protein maturation by protein folding', 'regulation of gene expression', 'RNA strand annealing activity', 'protein *O*-linked glycosylation', 'histone acetyltransferase complex'), energy production (for example, 'mitochondrial membrane'), stress response (for example, 'response to acid chemical'), immune response (for example, 'respiratory burst involved in inflammatory response', 'response to histamine', 'defense response') and metabolism (for example, '4-hydroxyproline metabolic process', 'positive regulation of protein metabolic process', 'negative regulation of gluconeogenesis').

Our candidate SNPs occurred in genomic regions within or proximate to several manually annotated ion transporter genes (Table 1, Supplementary Tables 5–8). Notably, the highest density of parallel candidate SNPs was found on scaffold 68 in a region that contains seven tandem paralogues of the Na^+/H^+ antiporter (*NHA*) (Fig. 2). Parallel candidate SNPs were found within or proximate to four paralogues of the Na^+/H^+ antiporter (*NHA* 3, 4, 5, 7), the α and β subunits of Na^+,K^+-ATPase (*NKA* α2, *NKA* β5) and one paralogue of the ammonium transporter Rh protein (*Rh* 2). Non-parallel candidate SNPs were found within or proximate to additional ion transporter genes, including multiple paralogs of *CA*, *NKCC*, *NHE*, *NKA* and *NHA* (Table 1).

SNPs with signatures of directional selection in the invading range are enriched for signatures of balancing selection in the native ranges. Our population genomic data support the hypothesis that a significant proportion of the genetic variation that responded to directional selection during freshwater invasions was maintained in the native populations by balancing selection. If the genetic variants that facilitated freshwater invasions were maintained by selection due to seasonally fluctuating salinity in the native ranges, this signature of long-term balancing selection should be detectable in the native saline populations. To test this prediction, we scanned the genomes of the saline populations for signatures of long-term balancing selection using two recently developed summary statistics, $\beta^{(2)}$ and Non-central Deviation (NCD2)[64–66], using all SNPs with a MAF > 0.05 in each population. These newer methods have been shown to be considerably more powerful and robust to non-equilibrium demographic histories than traditional tests, such as Tajima's *D*[67]. High $\beta^{(2)}$ scores indicate an excess of SNPs at similar frequencies, while low NCD2 scores indicate a build-up of SNPs near a specified intermediate frequency, both of which are potential consequences of long-term balancing selection. Both statistics also increase in significance with a deficit of substitutions relative to an outgroup. We first tested whether parallel and non-parallel candidate SNPs and 10-kb windows around our candidate SNPs were enriched for signatures of balancing selection relative to the whole genome in each native saline population. We then tested whether the SNPs with the strongest signatures of balancing selection in the native ranges were enriched for signatures of directional selection and association with salinity in the invading freshwater populations.

Notably, we found that candidate SNPs and windows with signatures of directional selection and association with salinity were enriched for signatures of long-term balancing selection in all native saline populations. Specifically, parallel candidate SNPs had significantly stronger signatures of balancing selection than the whole genome in four (using $\beta^{(2)}$) or three (using NCD2) of the native populations (Fig. 3, Supplementary Table 9, Extended Data Figs. 1 and 2). Mean $\beta^{(2)}$ scores were up to 3.65-fold higher in parallel candidate SNPs than the genome-wide average. On average, $\beta^{(2)}$ scores for SNPs in 10-kb windows around parallel candidate SNPs were 3.1-fold higher than the genome-wide average, and $\beta^{(2)}$ scores for SNPs in shared candidate windows were 2.35-fold higher than the genome-wide average. This enrichment of balancing selection in windows around parallel candidate SNPs and in shared windows was significant in all four native saline populations (Supplementary Table 9). Although the enrichment was less pronounced when using NCD2, there was a significant reduction in NCD2 scores (indicating stronger signatures of balancing selection) in all four native range populations (Supplementary Table 9). There was an average decrease in NCD2 score of 7.4% in 10-kb windows around parallel candidate SNPs and a decrease in NCD2 score of 4.9% in shared candidate windows.

In a complementary manner, SNPs with the strongest signatures of balancing selection in the native populations (the top 1% of the $\beta^{(2)}$ distributions and bottom 1% of the NCD2 distributions in each saline population) were significantly enriched for signatures of directional selection and association with salinity relative to the whole genome when considering SNPs shared by both clades (Supplementary Table 9). The one exception was the native range population in Baie de L'Isle Verte, where SNPs in the top 1% of the $\beta^{(2)}$ and NCD2 distributions were not enriched for signatures of directional selection, but were enriched for association with salinity (Supplementary Table 9). SNPs in the top 1% of $\beta^{(2)}$ scores and bottom 1% of NCD2 scores were, on average, 1.29 deciban units higher than the whole genome in the BayeScan 3 analysis. This value corresponds to a 1.35-fold BF increase in support for directional selection for these SNPs under balancing selection. Similarly, these SNPs had, on average, 1.34-fold higher support for association with salinity.

Overall, the strongest signatures of balancing selection were found in the parallel candidate SNPs rather than in shared windows or non-parallel candidate SNPs. Non-parallel candidate SNPs were also broadly enriched for signatures of balancing selection, with up to a 2.43-fold mean $\beta^{(2)}$ score increase in non-parallel candidate SNPs relative to the whole genome (Supplementary Table 10). Notably, parallel candidate SNPs showed stronger signatures of balancing selection than non-parallel candidate SNPs in two (using $\beta^{(2)}$)

Content courtesy of Springer Nature, terms of use apply. Rights reserved

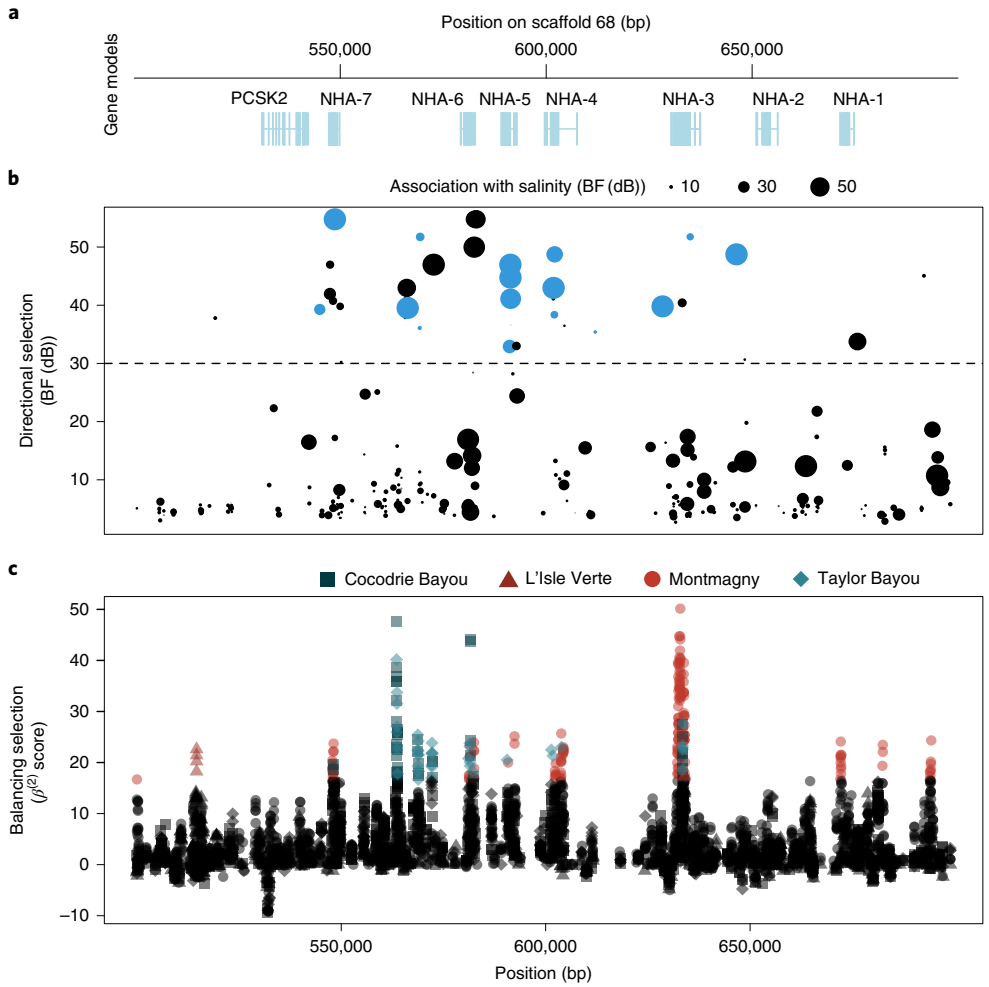


Fig. 2 | Signatures of directional and balancing selection at the *NHA* cluster in the *E. affinis* complex genome. a, Annotated genes on scaffold 68 of the *E. affinis* complex (Atlantic clade) draft genome[83], showing seven tandem paralogues of the *NHA* gene family. **b**, Support for directional selection was estimated with BayeScan 3 by comparing SNP frequencies among the native saline and freshwater invading populations of both the Atlantic and Gulf clades. The horizontal dotted line corresponds to the chosen significance level for directional selection. Blue dots correspond to the SNPs with significant support for parallel directional selection in the BayeScan 3 analysis. Large blue dots (BF>30, deciban units (dB)) above the dotted line are also significantly associated with salinity and are designated as the 'parallel candidate SNPs' (Supplementary Box 1). Black dots above the dotted line correspond to SNPs with support for directional selection in only one clade. The size of the dots corresponds to the level of BF support for association with salinity. **c**, Signatures of balancing selection (based on $\beta^{(2)}$ scores) in the four native saline populations (Fig. 1) from the Atlantic (red triangles and circles) and Gulf (green squares and diamonds) clades. All SNPs in each population for which $\beta^{(2)}$ scores were calculated are shown. Coloured data points represent SNPs in the top 1% of $\beta^{(2)}$ scores in each population.

or three (using NCD2) native range populations (Supplementary Table 11, section A, rows 1–4, 9–12). Similarly, 10-kb windows around parallel candidate SNPs had stronger signatures of balancing selection than 10-kb windows around non-parallel candidate SNPs in all four native range populations (Supplementary Table 11, section A, rows 5–8, 13–16). For example, in the native population in Cocodrie Bayou, $\beta^{(2)}$ scores for parallel candidate SNPs were 4.95-fold higher than non-parallel candidate SNPs and 1.34-fold higher than SNPs in shared windows (Supplementary Table 11, section A, row 3 and section C, row 3). These results suggest that many selected SNPs, comprising both parallel and non-parallel SNPs, were maintained by balancing selection in the native range. However, finding stronger signatures of balancing selection in the set of parallel candidate SNPs than in the set of non-parallel candidate SNPs supports the hypothesis that balancing selection specifically promotes parallel evolution by maintaining shared adaptive variation.

Parallel candidate SNPs exhibited significant signatures of balancing selection in the native range populations to a greater degree than expected using both neutral simulations and whole-genome distributions as null models to calculate significance. In fact, between 15.6% and 47.4% of parallel candidate SNPs fell in the top 5% of simulated balancing selection scores in the native range populations, and between 6.4% and 17.4% fell in the top 5% of the genome-wide distributions (Fig. 3c, Supplementary Table 12). These proportions of parallel candidate SNPs with significant signatures of balancing selection were broadly greater than the genome-wide expectations of around 15% based on the neutral simulation approach and 5% based on the whole-genome outlier approach (Fig. 3c, Supplementary Table 12). Considering the $\beta^{(2)}$ score, approximately 66.0% (using neutral simulations to calculate significance) and 12.5% (using the whole-genome distributions to calculate significance) of 10-kb windows around parallel candidate SNPs exhibited strong signatures of balancing selection in at least one native range population in both clades. These windows with signatures of balancing selection in both clades overlapped with several manually annotated ion transporter genes, including *NHA*

Content courtesy of Springer Nature, terms of use apply. Rights reserved

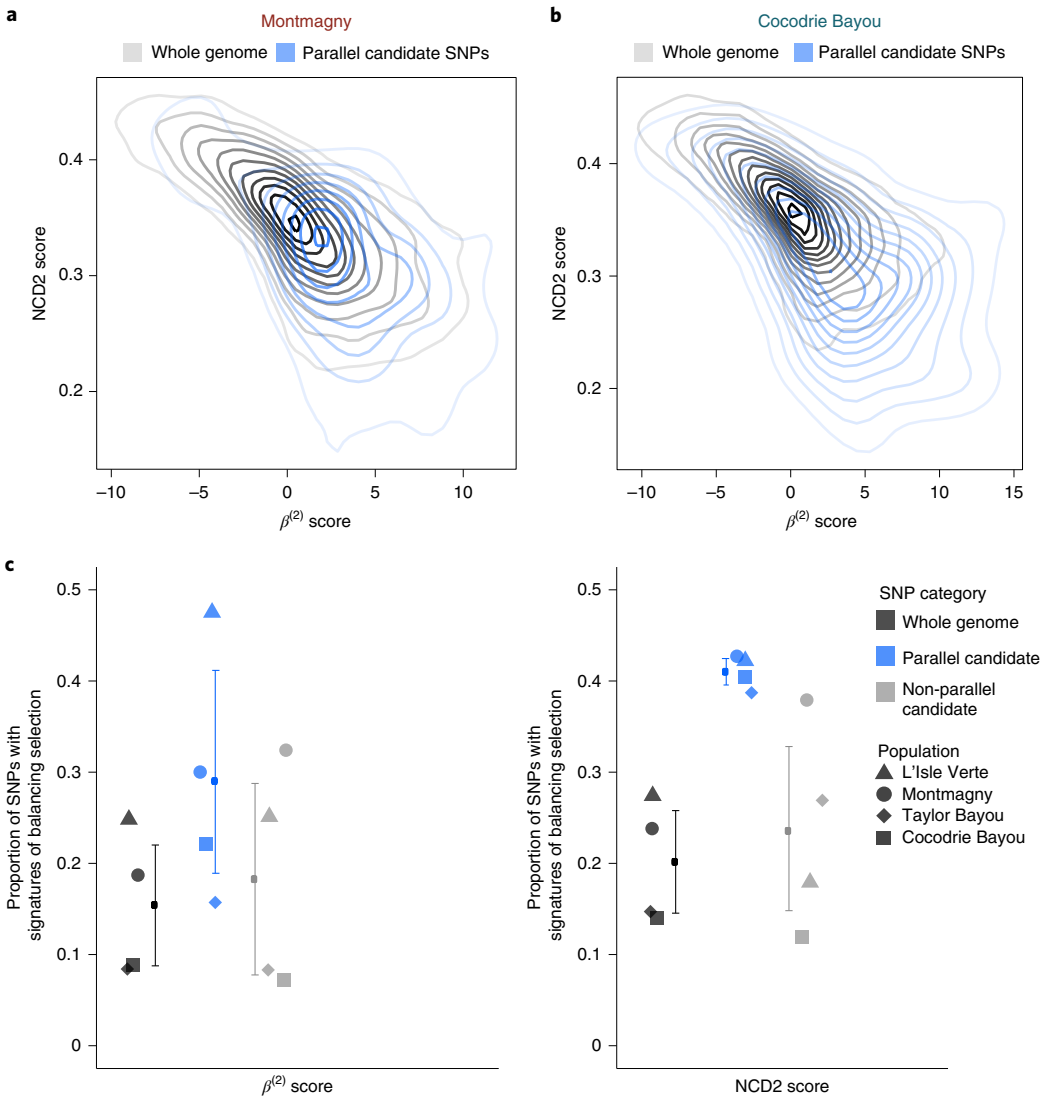


Fig. 3 | Loci with signatures of parallel directional selection and association with salinity are enriched for signatures of long-term balancing selection in the native ranges. Higher $\beta^{(2)}$ scores and lower NCD2 scores indicate stronger signatures of long-term balancing selection. On average, $\beta^{(2)}$ scores are higher and NCD2 scores are lower for parallel candidate SNPs (blue) relative to the rest of the genome (grey). **a**, $\beta^{(2)}$ and NCD2 score density plot for the native, saline Montmagny population (St Lawrence estuary, Atlantic clade). **b**, $\beta^{(2)}$ and NCD2 score density plot for the native, saline Cocodrie Bayou population (Gulf of Mexico, Gulf clade). **c**, Proportion of SNPs with strong signatures of balancing selection based on $\beta^{(2)}$ scores (left) (falling in the top 5% of the simulated null distribution) and NCD2 scores (right) (falling in the bottom 5% of the simulated null distribution). Data points are displayed for each of the four native saline populations (Fig. 1). A greater proportion of parallel candidate SNPs (blue data points; Supplementary Box 1) show signatures of balancing selection, relative to non-parallel candidate SNPs (grey points) and SNPs across the whole genome (black points). Error bars represent 95% confidence limits for the mean values of the four populations.

paralogues 6 and 7, and *NKA* β subunit, paralogue 5. Overall, these results indicate that loci with signatures of long-term balancing selection in the native ranges have a higher probability of responding to directional selection during invasions relative to loci without such signatures.

Discussion

A comprehensive understanding of the evolutionary genetic mechanisms that underlie successful invasions is key to the prediction, prevention and management of current and future biological invasions[5,9]. In this study, we present evidence for highly repeatable evolutionary genomic mechanisms that underlie rapid adaptation during invasions. Our study provides evidence, at the genomic level, that fluctuating habitats can promote successful invasions of novel habitats by maintaining the genetic variation required for rapid adaptation. We show that a substantial proportion of the genetic variants that underlie freshwater invasions are likely maintained by balancing selection in the native ranges, rather than by mutation-drift balance. In other words, the SNPs and genomic windows with signatures of balancing selection have higher probabilities of responding to directional selection during invasions than loci without signatures of balancing selection. Furthermore, our data provide evidence that balancing selection can enable parallel adaptation to repeatedly deploy the same standing variation preserved in the native range. The marked degree of parallelism at the genetic level appears to be the direct result of balancing selection maintaining a common pool of adaptive genetic variation in the native range populations.

Content courtesy of Springer Nature, terms of use apply. Rights reserved

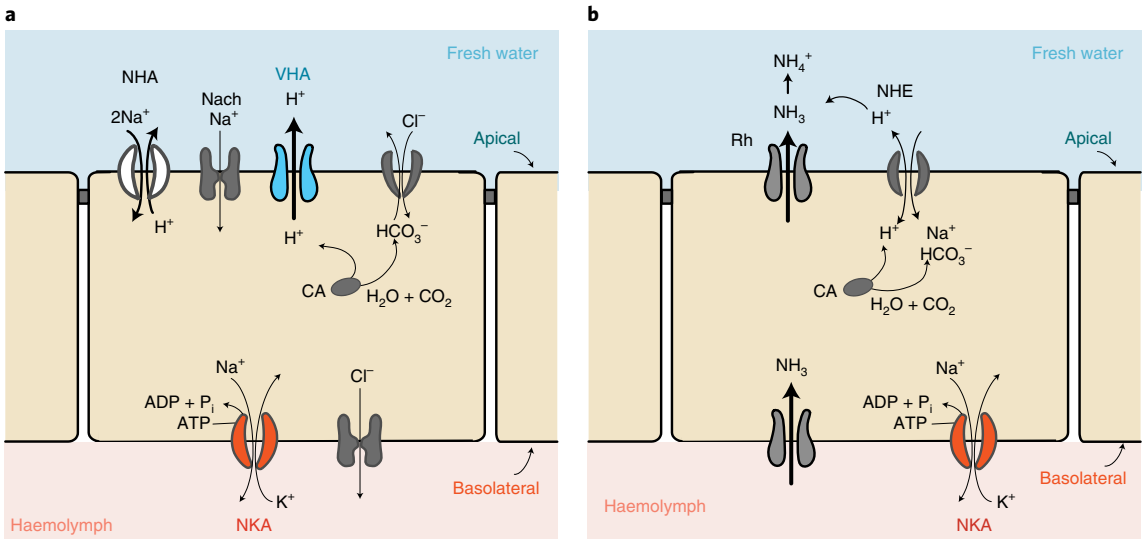


Fig. 4 | Alternative models of ion uptake from freshwater environments. a, Model 1. VHA generates an electrochemical gradient by pumping out protons, to facilitate uptake of Na^+ through an electrogenic Na^+ transporter (possibly NHA or Na^+ channel (Nach)). **b**, Model 2. An ammonia transporter Rh protein exports NH_3 out of the cell and this NH_3 reacts with H^+ to form NH_4^+. The reduced extracellular H^+ concentrations causes NHE to export H^+ in exchange for Na^+. These models are not comprehensive for all tissues or taxa and they are not mutually exclusive. Adapted from Lee et al.[63] and Dymowska et al.[98].

Despite the deep phylogenetic distance among the invasive populations in different clades, directional selection often acted on the same loci across invasion events. This parallelism was observed both at different SNPs within the same small genomic windows (possibly representing different SNPs at the same loci) and at exactly the same SNPs in both clades. Although a considerable number of loci had signatures of selection in only one clade, the signatures of selection on shared standing genetic variation was surprising given the large genetic divergences between the native range populations from different clades. Indeed, the detected proportion of candidate SNPs with support for signatures of parallel selection (42.5% when considering shared variation) was higher than several other well-documented cases of parallel evolution[68] (for example, 31% of SNPs underlie parallel high-altitude tolerance in humans[56] and 35% of genomic regions underlie parallel freshwater adaptation in the threespine stickleback[59]). In addition, our analysis found the strongest enrichment of balancing selection among parallel candidate SNPs rather than non-parallel candidate SNPs. Thus, we posit that many of those SNPs had remained segregating in the native populations for so long, making them available for natural selection in multiple populations, because they were maintained by balancing selection from the time preceding the split between the two clades.

According to classical theoretical studies, balancing selection as a result of temporally varying environments has been thought to be restricted in its ability to maintain polymorphisms through time, as the maladapted alleles would be removed from the population by negative selection[33–35,69]. Even as recent theoretical and empirical studies have suggested that fluctuating selection can maintain variation at many loci[70–72], evidence for this phenomenon has remained limited. Notably, native range populations of the *E. affinis* complex do exhibit several properties that would greatly expand the conditions under which balancing selection could maintain polymorphisms (Box 1). These features, along with the results presented here, make balancing selection a plausible mechanism for maintaining the observed key genetic variation required for adaptation during invasions. While we are not arguing that this mechanism is solely responsible for promoting successful invasions, this study provides strong empirical support that this mechanism is quite possible. Indeed, a large proportion of candidate SNPs did not have signatures of balancing selection in the native range, suggesting that neutral and rare variants also play an important role. Nevertheless, the significant enrichment for balancing selection among both parallel and non-parallel candidate loci does strongly suggest that balancing selection is a key mechanism promoting adaptation to novel habitats. Future studies should explore whether seasonal fluctuations in salinity do indeed induce correlated fluctuations in SNP frequencies for candidate loci, producing the genetic signatures of long-term balancing selection for critical traits.

Temporally fluctuating selection is a widespread phenomenon in general[73,74] and the copepod *E. affinis* complex is not unique among biological invaders in facing environmental heterogeneity in their native ranges[19]. For instance, invasive zebra mussel populations have been shown to originate from brackish estuaries marked by heterogeneity, rather than from the more stable ancient lakes[75,76]. Mechanistically, seasonal changes in dominance can greatly increase the extent of variation maintained by balancing selection, imposed by temporally fluctuating conditions, by protecting maladapted alleles from negative selection[70,77] (Box 1). While empirical evidence for environmentally dependent dominance has been demonstrated in the *E. affinis* complex[53] and in *Drosophila*[78], further work is needed to assess the pervasiveness of beneficial reversal of dominance in nature. Overlapping generations present a potentially prevalent mechanism that protects polymorphisms from negative selection during fluctuating conditions, given that dormancy, seed banks and diapause eggs are found commonly in nature, including in invasive populations[79]. As global climate change promises to modify the extent of spatial and temporal environmental heterogeneity, the ability to detect balancing selection in genomes may become a critical component of predicting future invasions, as well as responses to climate change.

The ability of temporally fluctuating habitats to promote success in introduced populations likely depends on the precise traits under balancing selection in the native habitats and how they relate to new selection pressures. In the present case, ion uptake appears to be a key physiological trait underlying rapid transitions from saline to freshwater habitats[61,62] (Table 1, Supplementary Table 4). However, fundamental mechanisms of ion uptake in fresh water continue to be debated, particularly the mechanism by which Na^+ is

Content courtesy of Springer Nature, terms of use apply. Rights reserved

transported into cells from very low external ionic conditions (see Fig. 4 for alternative hypotheses). Although our population genomic analyses implicate genes that are possibly involved in both of these models of ion uptake (for example, *NHA*, *NKA* or *Rh*), the strongest and most consistent signal of selection in both the Atlantic and Gulf clades occurred within the *NHA* paralogue cluster (Fig. 2). These results point to a key role of Na^+/H^+ antiporter and other ion-transporter paralogues in adaptive ion uptake in fresh water (Table 1, Fig. 4a). Future functional and phylogenetic studies within the Crustacea and across animals would reveal the functions of the ion-transporter paralogues and their general importance for adaptive evolution across salinity boundaries.

In conclusion, we have leveraged deep population genomic sampling of native and invasive populations of the *E. affinis* complex to uncover both parallel and clade-specific signatures of selection in response to freshwater invasions. We discovered that directional selection in the invasive populations frequently acted on shared standing polymorphisms with signatures of balancing selection in the native ranges. Our results point to a high degree of repeatability in the evolutionary processes that generate successful invaders, from the conditions in the native habitats to the selection response following introductions. Our findings suggest that balanced polymorphisms might often be an integral component of rapid adaptation to novel environments, contributing to a long-standing debate regarding the role of balancing selection in maintaining genetic variation within populations. Given that balancing selection may be more common in nature than previously thought, future studies of adaptation to new environments should consider the role of balancing selection in maintaining critical standing genetic variation upon which selection could act, especially for populations experiencing rapid environmental change.

Methods

Population sampling, raw read processing and SNP calling. The copepod *E. affinis* constitutes a cryptic species complex, comprising at least six geographically separated and genetically divergent clades distributed across the Northern Hemisphere[54,80]. Population sampling targeted nine wild populations of the *E. affinis* complex from two genetically distinct lineages (the Gulf and Atlantic clades)[49,54] (Fig. 1; Supplementary Table 1). Only the Atlantic clade has been named *E. carolleeae*[81], but is referred to as the Atlantic clade in this paper for consistency, given that the other clades of the *E. affinis* complex (for example, the Gulf clade) have not been separately named. In total, five invading freshwater (0–0.9 practical salinity units (PSU), which is approximately equal to parts per thousand salinity) populations and four native saline (>4 PSU) populations were used in this study. From each population, 100 adult individuals were sampled with an approximate 1:1 sex ratio. Individuals were pooled and whole-genome shotgun sequenced (that is, Pool-seq) using the Nextera DNA library preparation kit (Illumina). Libraries were sequenced on an Illumina HiSeq platform at the Institute for Genome Sciences, University of Maryland School of Medicine, generating an average of 179 million 100-bp read pairs per population. Raw reads were trimmed and filtered of adapter sequences, low-complexity sequences and low-quality ($Q<15$) bases using BBDuk in the BBTools package (https://jgi.doe.gov/data-and-tools/bbtools/). Processed reads were mapped to the repeat-masked[82] *E. affinis* complex (Atlantic clade) draft reference genome[83] using BWA-MEM v.0.7.17 (ref. [84]). Paired-end reads that did not align concordantly with BWA-MEM were aligned as single-end reads using NextGenMap v.0.5.5 to aid in the alignment of diverged sequences[85]. The combined read-mapping procedure achieved a mean mapping rate of 95.09 ± 1.42%. Duplicate reads were removed using Picard v.2.18.27 (http://broadinstitute.github.io/picard) and regions around insertions or deletions were realigned using GATK v.3.8 (ref. [86]). SAMtools v.1.3.1 was used to convert BAM files into mpileup format after removing low-quality alignments and bases ($Q<20$). Sites within 3 bp of an insertion or deletion were removed and the filtered mpileup was converted to sync format using PoPoolation2 (ref. [87]). The R package poolfstat v.1.0 (ref. [88]) was used to detect bi-allelic SNPs with a global MAF > 0.05, at least four reads were required for a base call, and a minimum of 20 and a maximum of 200 total read counts were required for all populations. In total, 6,635,765 SNPs passed these filters when considering all nine populations. A total of 7,565,621 and 5,323,780 SNPs were called for the Atlantic and Gulf clades, respectively.

Estimating population history. The software TreeMix v.1.13 was used to estimate a bifurcating population tree from bi-allelic SNP frequencies using windows of 1,000 SNPs in size[55]. Only SNPs with a MAF > 0.05 in both clades (that is, shared variants) were used for this analysis as the model does not consider new mutations. Node support was assessed using 100 bootstrap replicates.

To assess the potential importance of genetic drift during invasions, parameters indicative of effective population size ($\theta = 4N_e\mu$) were estimated for each population. PoPoolation[89] was used to estimate $\theta_{Watterson}$ and θ_π in 25-kb non-overlapping windows under an infinite sites model. These estimates used the correction for Pool-seq sampling, requiring at least four reads for a base call and a minimum of 20 and maximum of 200 total read counts for a SNP call. Phylogenetic generalized least squares regression was used to test for a relationship between genome-wide θ estimates and salinity. Phylogenetic generalized least squares analyses were performed in the R package phytools v.0.6 (ref. [90]) with the function pgls.SEy using the phylogeny estimated with TreeMix. Mean θ was regressed against square-root-transformed salinity, accounting for variance in θ for each population using a previously published method[91].

Detecting genomic signatures of parallel directional selection in the invading populations. BayeScan 2 (ref. [57]) was used to detect signatures of directional selection in each clade separately—that is, to identify the non-parallel candidate SNPs (Supplementary Box 1). The hierarchical version of BayeScan (that is, BayeScan 3 or BayeScanHierarchical[56]) was used to detect SNPs that display parallel signatures of directional selection in both the Atlantic and Gulf clades—that is, parallel candidate SNPs (Supplementary Box 1). For the detection of parallel candidate SNPs, the dataset was restricted to SNPs with a minimum MAF of 0.05 in both clades (that is, excluding alleles that are fixed in either clade), such that 366,781 SNPs remained.

In each BayeScan analysis, the Markov chain Monte Carlo (MCMC) chain was run three times independently with 20 pilot runs of 500 iterations each, a burn-in of 2,500 iterations and 1,000 samples collected with a thinning interval of 50 iterations. As parameter estimates were highly consistent across runs, parameter estimates were taken as the median of the three runs to remove potentially spurious estimates from any individual run. The prior odds for selection were set to 0.01 to match the default prior of the BayPass model[58] to facilitate a direct comparison of the BF results from the two models. For the BayeScan 3 analysis, the support for a model with selection versus neutrality was taken as the sum of the posterior probabilities of the three selection models (only the Atlantic clade, only the Gulf clade and parallel). These posterior probabilities for selection were then converted to BFs in deciban units ($10\log_{10}$(posterior odds/prior odds)). BF values greater than 30 deciban units were considered decisive evidence for selection, and the model of selection (only the Atlantic clade, only the Gulf clade or parallel) with the highest posterior probability was identified as the best fit model.
For the BayeScan 2 analyses, BF values were estimated from the α parameter and converted to deciban units as above. To analyse the large number of SNPs in the independent clade analyses, each dataset was subsampled into sets of around 100,000 random SNPs.

The BayPass v.2.1 package[58] was used to identify SNPs with frequencies that were significantly associated with salinity. Posterior distributions of the parameters of the Pool-seq version of the BayPass models were estimated for each SNP using the following MCMC procedure: 15 pilot runs of length 500 were run before a burn-in of 2,500 iterations and a sampling of 1,000 MCMC samples with a thinning interval of 25 iterations. To analyse the large number of SNPs in the independent clade analyses, each dataset was subsampled into sets of approximately 100,000 random SNPs. Parameter estimates were taken as the median of the three independent runs. To estimate the expected variance–covariance matrix for SNP frequencies, the posterior distributions of the parameters of the BayPass core model were estimated using the aforementioned MCMC procedure. To estimate the correlation between each SNP and salinity, posterior distributions of the parameters of the auxiliary model were estimated, considering salinity measured at the time of sample collection as the environmental covariate (Supplementary Table 1). As recommended[58], salinity was scaled so that the mean = 0 and variance = 1. The auxiliary δ parameter estimates the probability that the frequency of a SNP is correlated with a variable of interest. The δ parameter estimates were used to calculate BFs using the default prior distribution[58]. SNPs displaying both a significant (BF > 30) association with salinity and support for directional selection were considered to be candidate SNPs (Supplementary Box 1) that were putatively linked to targets of selection during freshwater invasions. Parallel candidate SNPs were also examined for aberrant coverage profiles indicating potential mapping artefacts that could result from copy-number variation or mapping biases due to the divergence between the two clades (Supplementary Information section II). Candidate SNPs with such signatures ($n = 2$) were removed from downstream analyses.

Approximate gene annotations were obtained by assigning SNPs to their closest gene model in the *E. affinis* complex (Atlantic clade) genome using BedTools v.2.28 (ref. [92]) and BEDOPS[93]. Putative ion-transport-related genes were manually annotated with the Web Apollo platform[94] using a combination of BLAST searches against the NCBI RefSeq and nr databases, i5K Arthropod Genome databases, FlyBase and FleaBase. To determine the specific paralogue clade identities for our candidate genes, phylogenies were constructed of candidate gene families across the Arthropoda as described previously[83].

Content courtesy of Springer Nature, terms of use apply. Rights reserved

GO enrichment tests were performed to detect functional groups that were enriched in each of our three categories of significant SNPs (Supplementary Box 1). We assigned GO terms to *E. affinis* complex gene models using significant ($E < 0.001$) BLAST hits in the Uniprot/Swissprot database. We used Gowinda[95] to detect GO terms that were significantly enriched in our candidate SNPs, using 100,000 simulations to assess significance. We assigned GO terms to SNPs falling within 10 kb of gene models (that is, including SNPs in potential regulatory regions) and counted only one SNP per coding region to account for linkage. As the tests were designed to be applied to SNP datasets, they were not performed on shared candidate windows. *P* values were corrected for multiple tests using the Benjamini–Hochberg procedure[96].

Detecting genomic signatures of balancing selection in the native range populations. To test whether the SNPs under directional selection in the freshwater invading populations were maintained by balancing selection in the native range saline populations, genomic scans for balancing selection were performed for each of the four native range populations using the $\beta^{(2)}$ and NCD2 statistics. In contrast to traditional tests, such as Tajima's *D* and the Hudson–Kreitman–Aguade test, these methods explicitly consider the expected allele frequency distribution under long-term balancing selection. High $\beta^{(2)}$ scores indicate an excess of SNPs at similar frequencies[64,66], while low NCD2 scores indicate a build-up of SNPs near a specified intermediate frequency[62]. Both statistics also increase in significance with a deficit of substitutions relative to an outgroup.

$\beta^{(2)}$ and NCD2 scores were calculated for SNPs with a MAF > 0.15 to reduce false positives[64–66]. This MAF threshold was determined previously[64], based on simulations suggesting that balanced SNPs were unlikely to achieve equilibrium frequencies <0.15 due to their high probability of drifting out of the population. A MAF filter of 0.05, evaluated for each population separately, was applied to SNPs in the windows used to calculate the scores. SNP frequencies were polarized and substitutions were called using the alternate clade as the outgroup. A window size of 500 bp was used based on the estimated recombination rate of the copepod *Tigriopus californicus*[97]. The NCD2 statistic was developed to calculate scores for genomic windows, while $\beta^{(2)}$ calculates scores for individual SNPs. To facilitate comparison between the two statistics, a custom Python script was used to calculate a modified NCD2 statistic for every SNP (rather than genomic window) in the saline population genomes using windows of 500 bp around each SNP and considering a target frequency of 0.5. Although this target frequency was chosen arbitrarily, NCD2 scores have been shown to be robust to the choice of target frequency[65].

To estimate a null distribution of $\beta^{(2)}$ and NCD2 scores under a model with only genetic drift, we calculated $\beta^{(2)}$ and NCD2 scores for SNPs simulated under the parameters of the neutral variance–covariance matrix (that is, population history). Simulated read counts (that is, SNP frequencies) were generated using the R function simulate.baypass included in the BayPass v.2.1 package using the core BayPass model parameters inferred from the full SNP dataset. The simulated SNPs were placed in the empirical SNP positions along the genome to calculate the null distribution of $\beta^{(2)}$ and NCD2 scores for each native range population. Empirical values were then compared with these null distributions to determine deviation from the null model as a measure of significance. As our simulation approach ultimately assumed that a large proportion of the genome was subject to balancing selection, we also used more conservative empirical cut-offs based on the whole-genome distributions.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Raw sequence data and aligned reads have been deposited in the NCBI Sequence Read Archive under BioProject ID PRJNA610547.

Code availablility

Custom scripts used throughout this analysis are available online at https://github.com/TheDBStern/NEE2020.

Received: 27 August 2019; Accepted: 14 April 2020;
Published online: 22 June 2020

References

1. Paini, D. R. et al. Global threat to agriculture from invasive species. *Proc. Natl Acad. Sci. USA* **113**, 7575–7579 (2016).
2. Bradshaw, C. J. A. et al. Massive yet grossly underestimated global costs of invasive insects. *Nat. Commun.* **7**, 12986 (2016).
3. Bellard, C. et al. Will climate change promote future invasions? *Glob. Change Biol.* **19**, 3740–3748 (2013).
4. Hellmann, J. J., Byers, J. E., Bierwagen, B. G. & Dukes, J. S. Five potential consequences of climate change for invasive species. *Conserv. Biol.* **22**, 534–543 (2008).
5. Chown, S. L. et al. Biological invasions, climate change and genomics. *Evol. Appl.* **8**, 23–46 (2015).
6. Rahel, F. J. & Olden, J. D. Assessing the effects of climate change on aquatic invasive species. *Conserv. Biol.* **22**, 521–533 (2008).
7. Chapman, D. S. et al. Modelling the introduction and spread of non-native species: international trade and climate change drive ragweed invasion. *Glob. Change Biol.* **22**, 3067–3079 (2016).
8. Dam, H. G. Evolutionary adaptation of marine zooplankton to global change. *Annu. Rev. Mar. Sci.* **5**, 349–370 (2013).
9. Blackburn, T. M. et al. A proposed unified framework for biological invasions. *Trends Ecol. Evol.* **26**, 333–339 (2011).
10. Lee, C. E. Evolutionary genetics of invasive species. *Trends Ecol. Evol.* **17**, 386–391 (2002).
11. Lee, C. E. in *Encyclopedia of Biological Invasions* (eds Simberloff, D. & Rejmanek, M.) 215–222 (Univ. California Press 2010).
12. Colautti, R. I., Alexander, J. M., Dlugosch, K. M., Keller, S. R. & Sultan, S. E. Invasions and extinctions through the looking glass of evolutionary ecology. *Phil. Trans. R. Soc. B* **372**, 20160031 (2017).
13. Williamson, M. & Fitter, A. The varying success of invaders. *Ecology* **77**, 1661–1666 (1996).
14. Hayes, K. R. & Barry, S. C. Are there any consistent predictors of invasion success? *Biol. Invasions* **10**, 483–506 (2008).
15. Bock, D. G. et al. What we still don't know about invasion genetics. *Mol. Ecol.* **24**, 2277–2297 (2015).
16. Kolar, C. S. & Lodge, D. M. Progress in invasion biology: predicting invaders. *Trends Ecol. Evol.* **16**, 199–204 (2001).
17. Lee, C. E. & Bell, M. A. Causes and consequences of recent freshwater invasions by saltwater animals. *Trends Ecol. Evol.* **14**, 284–288 (1999).
18. Casties, I., Seebens, H. & Briski, E. Importance of geographic origin for invasion success: a case study of the North and Baltic seas versus the Great Lakes–St. Lawrence River region. *Ecol. Evol.* **6**, 8318–8329 (2016).
19. Lee, C. E. & Gelembiuk, G. W. Evolutionary origins of invasive populations. *Evol. Appl.* **1**, 427–448 (2008).
20. Winkler, G., Dodson, J. J. & Lee, C. E. Heterogeneity within the native range: population genetic analyses of sympatric invasive and noninvasive clades of the freshwater invading copepod *Eurytemora affinis*. *Mol. Ecol.* **17**, 415–430 (2008).
21. Barrett, R. D. H. & Schluter, D. Adaptation from standing genetic variation. *Trends Ecol. Evol.* **23**, 38–44 (2008).
22. Lee, C. E. Evolutionary mechanisms of habitat invasions, using the copepod *Eurytemora affinis* as a model system. *Evol. Appl.* **9**, 248–270 (2016).
23. Peischl, S. & Excoffier, L. Expansion load: recessive mutations and the role of standing genetic variation. *Mol. Ecol.* **24**, 2084–2094 (2015).
24. Messer, P. W., Ellner, S. P. & Hairston, N. G. Can population genetics adapt to rapid evolution? *Trends Genet.* **32**, 408–418 (2016).
25. Lewontin, R. *The Genetic Basis of Evolutionary Change* (Columbia Univ. Press, 1974).
26. Crow, J. F. Muller, Dobzhansky, and overdominance. *J. Hist. Biol.* **20**, 351–380 (1987).
27. Beatty, J. Weighing the risks: stalemate in the classical/balance controversy. *J. Hist. Biol.* **20**, 289–319 (1987).
28. Asthana, S., Schmidt, S. & Sunyaev, S. A limited role for balancing selection. *Trends Genet.* **21**, 30–32 (2005).
29. Muller, H. J. Our load of mutations. *Am. J. Hum. Genet.* **2**, 111–176 (1950).
30. Dobzhansky, T. A review of some fundamental concepts and problems of population genetics. *Cold Spring Harb. Symp. Quant. Biol.* **20**, 1–15 (1955).
31. Hedrick, P. W. Genetic variation in a heterogeneous environment. I. Temporal heterogeneity and absolute dominance model. *Genetics* **78**, 757–770 (1974).
32. Dempster, E. R. Maintenance of genetic heterogeneity. *Cold Spring Harb. Symp. Quant. Biol.* **20**, 25–32 (1955).
33. Haldane, J. B. S. & Jayakar, S. D. Polymorphism due to selection of varying direction. *J. Genet.* **58**, 237–242 (1963).
34. Gillespie, J. Polymorphism in random environments. *Theor. Popul. Biol.* **4**, 193–195 (1973).
35. Felsenstein, J. Theoretical population genetics of variable selection and migration. *Annu. Rev. Genet.* **10**, 253–280 (1976).
36. Holt, R. D., Barfield, M. & Gomulkiewicz, R. Temporal variation can facilitate niche evolution in harsh sink environments. *Am. Nat.* **164**, 187–200 (2004).
37. Huang, Y., Tran, I. & Agrawal, A. F. Does genetic variation maintained by environmental heterogeneity facilitate adaptation to novel selection? *Am. Nat.* **188**, 27–37 (2016).
38. de Filippo, C. et al. Recent selection changes in human genes under long-term balancing selection. *Mol. Biol. Evol.* **33**, 1435–1447 (2016).
39. Hermisson, J. & Pennings, P. S. Soft sweeps: molecular population genetics of adaptation from standing genetic variation. *Genetics* **169**, 2335–2352 (2005).

Content courtesy of Springer Nature, terms of use apply. Rights reserved

40. Hermisson, J. & Pennings, P. S. Soft sweeps and beyond: understanding the patterns and probabilities of selection footprints under rapid adaptation. *Methods Ecol. Evol.* **8**, 700–716 (2017).
41. Jensen, J. D. On the unfounded enthusiasm for soft selective sweeps. *Nat. Commun.* **5**, 6281 (2014).
42. Ralph, P. L. & Coop, G. The role of standing variation in geographic convergent adaptation. *Am. Nat.* **186**, S5–S23 (2015).
43. Brennan, R. S., Garrett, A. D., Huber, K. E., Hargarten, H. & Pespeni, M. H. Rare genetic variation and balanced polymorphisms are important for survival in global change conditions. *Proc. R. Soc. B* **286**, 20190943 (2019).
44. Mallard, F., Nolte, V., Tobler, R., Kapun, M. & Schlotterer, C. A simple genetic basis of adaptation to a novel thermal environment results in complex metabolic rewiring in *Drosophila*. *Genome Biol.* **19**, 119 (2018).
45. Kelly, J. K. & Hughes, K. A. Pervasive linked selection and intermediate-frequency alleles are implicated in an evolve-and-resequencing experiment of *Drosophila simulans*. *Genetics* **211**, 943–961 (2019).
46. Stern, D. L. The genetic causes of convergent evolution. *Nat. Rev. Genet.* **14**, 751–764 (2013).
47. Lee, K. M. & Coop, G. Population genomics perspectives on convergent adaptation. *Phil. Trans. R. Soc. B* **374**, 20180236 (2019).
48. Waldvogel, A. M. et al. Evolutionary genomics can improve prediction of species' responses to climate change. *Evol. Lett.* **4**, 4–18 (2019).
49. Lee, C. E. Rapid and repeated invasions of fresh water by the copepod *Eurytemora affinis*. *Evolution* **53**, 1423–1434 (1999).
50. Katajisto, T. Copepod eggs survive a decade in the sediments of the Baltic Sea. *Hydrobiologia* **320**, 153–159 (1996).
51. Ban, S. & Minoda, T. Hatching of diapause eggs of *Eurytemora affinis* (Copepoda: Calanoida) collected from lake-bottom sediments. *J. Crustac. Biol.* **12**, 51–56 (1992).
52. Glippa, O., Denis, L., Lesourd, S. & Souissi, S. Seasonal fluctuations of the copepod resting egg bank in the middle Seine estuary, France: impact on the nauplii recruitment. *Estuar. Coast. Mar. Sci.* **142**, 60–67 (2014).
53. Posavi, M., Gelembiuk, G. W., Larget, B. & Lee, C. E. Testing for beneficial reversal of dominance during salinity shifts in the invasive copepod *Eurytemora affinis*, and implications for the maintenance of genetic variation. *Evolution* **68**, 3166–3183 (2014).
54. Lee, C. E. Global phylogeography of a cryptic copepod species complex and reproductive isolation between genetically proximate "populations". *Evolution* **54**, 2014–2027 (2000).
55. Pickrell, J. K. & Pritchard, J. K. Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet.* **8**, e1002967 (2012).
56. Foll, M., Gaggiotti, O. E., Daub, J. T., Vatsiou, A. & Excoffier, L. Widespread signals of convergent adaptation to high altitude in Asia and America. *Am. J. Hum. Genet.* **95**, 394–407 (2014).
57. Foll, M. & Gaggiotti, O. A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. *Genetics* **180**, 977–993 (2008).
58. Gautier, M. Genome-wide scan for adaptive divergence and association with population-specific covariates. *Genetics* **201**, 1555–1579 (2015).
59. Jones, F. C. et al. The genomic basis of adaptive evolution in threespine sticklebacks. *Nature* **484**, 55–61 (2012).
60. Alberto, F. J. et al. Convergent genomic signatures of domestication in sheep and goats. *Nat. Commun.* **9**, 813 (2018).
61. Gerber, L. et al. The legs have it: in situ expression of ion transporters V-Type H^+ ATPase and Na^+/K^+-ATPase in osmoregulatory leg organs of the invading copepod *Eurytemora affinis*. *Physiol. Biochem. Zool.* **89**, 233–250 (2016).
62. Johnson, K. E., Perreau, L., Charmantier, G., Charmantier-Daures, M. & Lee, C. E. Without gills: localization of osmoregulatory function in the copepod *Eurytemora affinis*. *Physiol. Biochem. Zool.* **87**, 310–324 (2014).
63. Lee, C. E., Kiergaard, M., Gelembiuk, G. W., Eads, B. D. & Posavi, M. Pumping ions: rapid parallel evolution of ionic regulation following habitat invasions. *Evolution* **65**, 2229–2244 (2011).
64. Siewert, K. M. & Voight, B. F. Detecting long-term balancing selection using allele frequency correlation. *Mol. Biol. Evol.* **34**, 2996–3005 (2017).
65. Bitarello, B. D. et al. Signatures of long-term balancing selection in human genomes. *Genome Biol. Evol.* **10**, 939–955 (2018).
66. Siewert, K. M. & Voight, B. F. BetaScan2: standardized statistics to detect balancing selection utilizing substitution data. *Genome Biol. Evol.* **12**, 3873–3877 (2020).
67. Tajima, F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585–595 (1989).
68. Bolnick, D. I., Barrett, R. D. H., Oke, K. B., Rennison, D. J. & Stuart, Y. E. (Non)parallel evolution. *Ann. Rev. Ecol. Evol. Syst.* **49**, 303–330 (2018).
69. Hedrick, P. W., Ginevan, M. E. & Ewing, E. P. Genetic polymorphism in heterogeneous environments. *Ann. Rev. Ecol. Syst.* **7**, 1–32 (1976).
70. Wittmann, M. J., Bergland, A. O., Feldman, M. W., Schmidt, P. S. & Petrov, D. A. Seasonally fluctuating selection can maintain polymorphism at many loci via segregation lift. *Proc. Natl Acad. Sci. USA* **114**, E9932–E9941 (2017).
71. Bergland, A. O., Behrman, E. L., O'Brien, K. R., Schmidt, P. S. & Petrov, D. A. Genomic evidence of rapid and stable adaptive oscillations over seasonal time scales in *Drosophila*. *PLoS Genet.* **10**, e1004775 (2014).
72. Troth, A., Puzey, J. R., Kim, R. S., Willis, J. H. & Kelly, J. K. Selective trade-offs maintain alleles underpinning complex trait variation in plants. *Science* **361**, 475–478 (2018).
73. Siepielski, A. M., DiBattista, J. D. & Carlson, S. M. It's about time: the temporal dynamics of phenotypic selection in the wild. *Ecol. Lett.* **12**, 1261–1276 (2009).
74. Thurman, T. J. & Barrett, R. D. H. The genetic consequences of selection in natural populations. *Mol. Ecol.* **25**, 1429–1448 (2016).
75. Gelembiuk, G. W., May, G. E. & Lee, C. E. Phylogeography and systematics of zebra mussels and related species. *Mol. Ecol.* **15**, 1033–1050 (2006).
76. May, G. E., Gelembiuk, G. W., Panov, V. E., Orlova, M. I. & Lee, C. E. Molecular ecology of zebra mussel invasions. *Mol. Ecol.* **15**, 1021–1031 (2006).
77. Bertram, J. & Masel, J. Different mechanisms drive the maintenance of polymorphism at loci subject to strong versus weak fluctuating selection. *Evolution* **73**, 883–896 (2019).
78. Chen, J., Nolte, V. & Schlotterer, C. Temperature stress mediates decanalization and dominance of gene expression in *Drosophila melanogaster*. *PLoS Genet.* **11**, e1004883 (2015).
79. Panov, V. E., Krylov, P. I. & Riccardi, N. Role of diapause in dispersal and invasion success by aquatic invertebrates. *J. Limnol.* **63**, 56–69 (2004).
80. Lee, C. E. & Frost, B. W. Morphological stasis in the *Eurytemora affinis* species complex (Copepoda: Temoridae). *Hydrobiologia* **480**, 111–128 (2002).
81. Alekseev, V. R. & Souissi, A. A new species within the *Eurytemora affinis* complex (Copepoda: Calanoida) from the Atlantic Coast of USA, with observations on eight morphologically different European populations. *Zootaxa* **2767**, 41–56 (2011).
82. Smit, A. F. A., Hubley, R. & Green, P. RepeatMasker Open-4.0 (2013–2015); http://www.repeatmasker.org
83. Eyun, S. I. et al. Evolutionary history of chemosensory-related gene families across the Arthropoda. *Mol. Biol. Evol.* **34**, 1838–1862 (2017).
84. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
85. Sedlazeck, F. J., Rescheneder, P. & von Haeseler, A. NextGenMap: fast and accurate read mapping in highly polymorphic genomes. *Bioinformatics* **29**, 2790–2791 (2013).
86. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
87. Kofler, R., Pandey, R. V. & Schlotterer, C. PoPoolation2: identifying differentiation between populations using sequencing of pooled DNA samples (Pool-Seq). *Bioinformatics* **27**, 3435–3436 (2011).
88. Hivert, V., Leblois, R., Petit, E. J., Gautier, M. & Vitalis, R. Measuring genetic differentiation from Pool-Seq data. *Genetics* **210**, 315–330 (2018).
89. Kofler, R. et al. PoPoolation: a toolbox for population genetic analysis of next generation sequencing data from pooled individuals. *PLoS ONE* **6**, e0015925 (2011).
90. Revell, L. J. phytools: an R package for phylogenetic comparative biology (and other things). *Methods Ecol. Evol.* **3**, 217–223 (2012).
91. Ives, A. R., Midford, P. E. & Garland, T. Within-species variation and measurement error in phylogenetic comparative methods. *Syst. Biol.* **56**, 252–270 (2007).
92. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
93. Neph, S. et al. BEDOPS: high-performance genomic feature operations. *Bioinformatics* **28**, 1919–1920 (2012).
94. Lee, E. et al. Web Apollo: a web-based genomic annotation editing platform. *Genome Biol.* **14**, R93 (2013).
95. Kofler, R. & Schlotterer, C. Gowinda: unbiased analysis of gene set enrichment for genome-wide association studies. *Bioinformatics* **28**, 2084–2085 (2012).
96. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate—a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* **57**, 289–300 (1995).
97. Foley, B. R.et al. A gene-based SNP resource and linkage map for the copepod *Tigriopus californicus*. *BMC Genom.* **12**, 568 (2011).
98. Dymowska, A. K., Hwang, P. P. & Goss, G. G. Structure and function of ionocytes in the freshwater fish gill. *Respir. Physiol. Neurobiol.* **184**, 282–292 (2012).
99. Lee, C. E. & Petersen, C. H. Effects of developmental acclimation on adult salinity tolerance in the freshwater-invading copepod *Eurytemora affinis*. *Physiol. Biochem. Zool.* **76**, 296–301 (2003).
100. Lee, C. E., Remfert, J. L. & Chang, Y. M. Response to selection and evolvability of invasive populations. *Genetica* **129**, 179–192 (2007).
101. Lee, C. E., Remfert, J. L. & Gelembiuk, G. W. Evolution of physiological tolerance and performance during freshwater invasions. *Integr. Comp. Biol.* **43**, 439–449 (2003).

Content courtesy of Springer Nature, terms of use apply. Rights reserved

102. Ellner, S. & Sasaki, A. Patterns of genetic polymorphism maintained by fluctuating selection with overlapping generations. *Theor. Popul. Biol.* **50**, 31–65 (1996).
103. Turelli, M. & Barton, N. H. Polygenic variation maintained by balancing selection: pleiotropy, sex-dependent allelic effects and GxE interactions. *Genetics* **166**, 1053–1079 (2004).
104. Turelli, M., Schemske, D. W. & Bierzychudek, P. Stable two-allele polymorphisms maintained by fluctuating fitnesses and seed banks: protecting the blues in *Linanthus parryae*. *Evolution* **55**, 1283–1298 (2001).
105. Ellner, S. & Hairston, N. G. Role of overlapping generations in maintaining genetic variation in a fluctuating environment. *Am. Nat.* **143**, 403–417 (1994).
106. Wright, S. Physiological and evolutionary theories of dominance. *Am. Nat.* **68**, 24–53 (1934).
107. Curtsinger, J. W., Service, P. M. & Prout, T. Antagonistic pleiotropy, reversal of dominance, and genetic polymorphism. *Am. Nat.* **144**, 210–228 (1994).
108. Gulisija, D., Kim, Y. & Plotkin, J. B. Phenotypic plasticity promotes balanced polymorphism in periodic environments by a genomic storage effect. *Genetics* **202**, 1437–1448 (2016).
109. Gulisija, D. & Plotkin, J. B. Phenotypic plasticity promotes recombination and gene clustering in periodic environments. *Nat. Commun.* **8**, 2041 (2017).

Acknowledgements

This research was funded by National Science Foundation grants OCE-1046372 and OCE-1658517 to C.E.L. and the Michael Guyer Postdoctoral Fellowship to D.B.S.

We thank J. Pool, S. Schoville and N. Sharpe for comments on the manuscript. The samples used in this project were collected by M. Bontrager. Computation was performed using the computational resources and assistance of the UW-Madison Center for High Throughput Computing (CHTC) in the Department of Computer Sciences.

Author contributions

C.E.L. contributed to the design of the study and data collection. D.B.S. contributed to the conception and execution of the analyses. Both authors contributed to the writing and approval of the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41559-020-1201-y.

Supplementary information is available for this paper at https://doi.org/10.1038/s41559-020-1201-y.

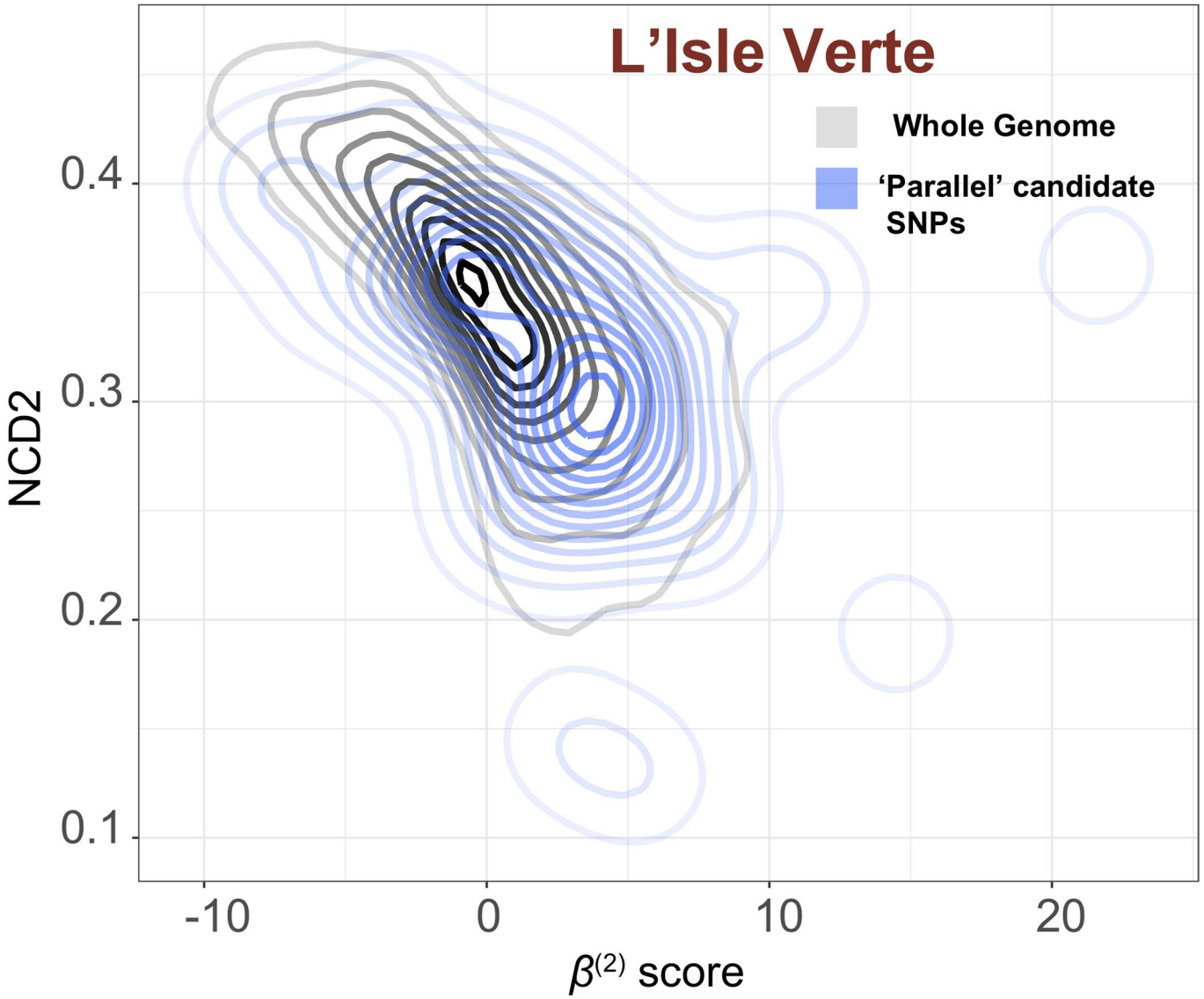
Correspondence and requests for materials should be addressed to C.E.L.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

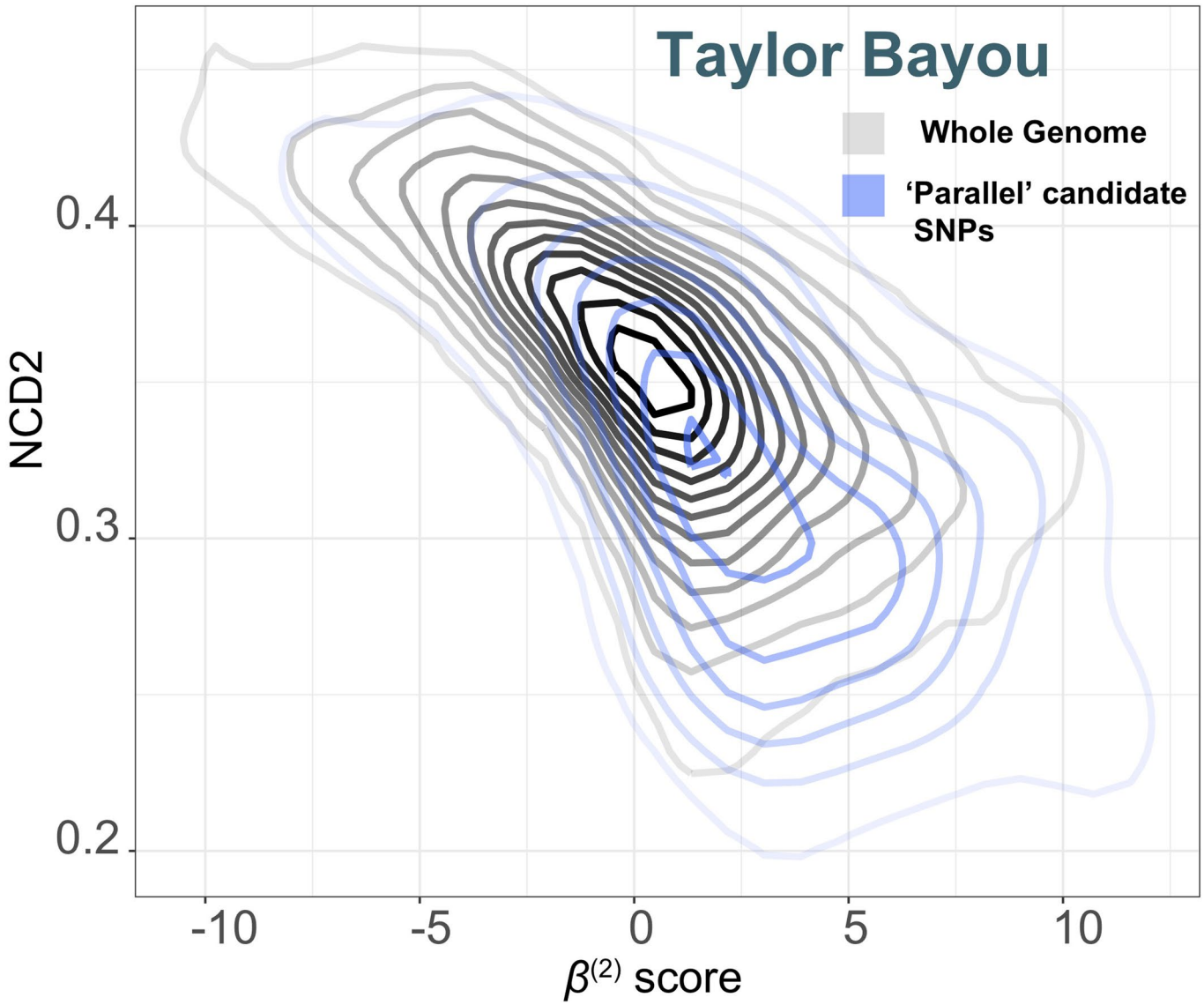
© The Author(s), under exclusive licence to Springer Nature Limited 2020

Content courtesy of Springer Nature, terms of use apply. Rights reserved



Extended Data Fig. 1 | $\beta^{(2)}$ and NCD2 score density plot for the native, saline Baie de L'Isle Verte population (St. Lawrence drainage, Atlantic clade). Higher $\beta^{(2)}$ scores and lower NCD2 scores signify stronger signatures of long-term balancing selection. $\beta^{(2)}$ scores are higher and NCD2 scores are lower on average for 'parallel' candidate SNPs (blue) relative to the rest of the genome (grey).

Content courtesy of Springer Nature, terms of use apply. Rights reserved



Extended Data Fig. 2 | $\beta^{(2)}$ and NCD2 score density plot for the native, saline Taylor Bayou (Gulf of Mexico, Gulf clade). Higher $\beta^{(2)}$ scores and lower NCD2 scores signify stronger signatures of long-term balancing selection. $\beta^{(2)}$ scores are higher and NCD2 scores are lower on average for 'parallel' candidate SNPs (blue) relative to the rest of the genome (grey).

Content courtesy of Springer Nature, terms of use apply. Rights reserved

Supplementary Information

Supplementary Box 1. Definitions of terms relevant to the detection of genomic signatures of natural selection

Term	Definition	Signal measured in this study	Method Used
Signatures of selection (Definitions are general)			
Directional selection	Natural selection favoring one allele over its alternate. We use the term directional selection because our data cannot inform whether the novel environment in the invaded range imposes positive or negative selection on a specific allele relative to the native range.	SNP frequency difference (F_{ST}) between native saline and invasive freshwater populations. Significant differentiation could indicate the action of natural selection favoring different alleles in native and saline populations.	BayeScan 2 or BayeScan 3
Association with salinity	The extent to which a shift in allele (SNP) frequency is correlated with changes in salinity. Such a correlation would suggest a functional relationship between the rise of particular alleles and freshwater adaptation.	Correlated changes in SNP frequency and salinity. Significant correlations could indicate the action of natural selection favoring different alleles in native saline and freshwater invasive populations.	BayPass 2.1
Balancing selection	Selection favoring more than one allele at a locus, resulting in the maintenance of genetic variation within a population.	Genomic windows with highly correlated SNP frequencies (β) or an excess of intermediate frequency SNPs (NCD2).	BetaScan2, NCD2
Candidate loci associated with adaptation during freshwater invasions (Definitions are specific to this study)			
'Parallel' candidate SNP	SNP exhibiting a frequency shift associated with freshwater invasions in both clades, representing selection on the same SNP in both clades.	SNP that is variable (MAF > 0.05) in both the Atlantic and Gulf clades with a significant signature of directional selection, support for the 'parallel' BayeScan 3 model, and association with salinity, analyzing all sampled populations jointly.	Bayescan 3 and BayPass 2.1
'Non-parallel' candidate SNP	SNP exhibiting a frequency shift associated with freshwater invasions in either clade alone.	SNP detected in either clade alone with a significant signature of directional selection and association with salinity when analyzing each clade separately.	Bayescan 2 and BayPass 2.1
'Shared' candidate window	Small genomic window with signatures of directional selection and association with salinity in both clades, but at different SNPs in each clade.	Overlapping BayeScan 2 significant and BayPass significant 10-kb windows between the Atlantic and Gulf clades.	Bayescan 2, BayPass 2.1, Bedtools

Supplementary Materials and Methods

I. Assessing the genome-wide significance of repeated evolution

a. Parallel frequency shifts in shared SNPs

We sought to detect parallel signatures of selection on shared standing genetic variation between the two clades, performing genome scans on SNPs with a MAF > 0.05 in both clades. To assess the probability of detecting parallel frequency shifts due to drift alone, we simulated neutral SNP frequencies for the nine populations under the inferred variance-covariance matrix, a parameter informative about the population history[1], and retained only SNPs with a MAF > 0.05 in both clades. The variance-covariance matrix was

inferred using the BayPass core model as described in the main text and the simulations were performed using the R function *simulate.baypass* included in the BayPass v2.1 package. The simulated SNPs should only show signatures of parallel directional selection by chance, i.e. due to non-selective processes, and therefore generate a null distribution of the expected degree of parallelism under drift. We then tested whether the number of SNPs showing significant signatures of parallel selection in the empirical dataset was greater than in the simulated neutral dataset using the same significance thresholds. We found that, among SNPs with significant signatures of directional selection with a MAF > 0.05 in both clades, the proportion of SNPs with the highest support for the parallel selection model was 1.73-fold greater than what was expected by drift alone (Supplementary Table 3). Additionally, we found that the empirical number of significant SNPs was significantly greater than the number in the neutral simulations examining sets of SNPs that show (1) association with salinity, (2) parallel selection, and (3) both association with salinity and parallel selection (Supplementary Table 3). The empirical dataset displayed 29.2 times the number of SNPs than the simulated dataset with both a signature of parallel directional selection and association with salinity (i.e. 'parallel' candidate SNPs; Supplementary Table 3). Thus, the observed degree of parallel SNP frequency shifts associated with freshwater invasions was significantly greater than the expectation from drift alone.

b. Shared windows between Atlantic and Gulf clade analyses, i.e. 'shared' candidate windows

Given the apparent divergence between the Atlantic and Gulf clades, we sought to detect signatures of selection at the same small genomic windows (10-kb), representing repeated selection at the same loci but potentially different SNPs. The window size of 10 kb was chosen capture selection targets on the same 'gene' and surrounding genomic region, given an average protein coding 'gene' size of ~8.5 kb in the *E. affinis* genome. As described in the main text, we performed separate SNP calling and genome scans for signatures of directional selection (BayeScan 2[2]) and association with salinity (BayPass v.2.1[3]) in each of the Atlantic and Gulf clades. The SNP calling pipeline resulted in 7,565,621 and 5,323,780 SNPs in Atlantic and Gulf clades, respectively, with a minor allele frequency (MAF) > 0.05. Given the small

percentage shared SNPs with a MAF > 0.05 in both clades (N=366,781), we took a window-based approach to detect shared genomic regions associated with directional selection and association with salinity.

While there was significant overlap between Atlantic and Gulf clade windows when considering SNPs that were both BayeScan and BayPass significant (N=29; Fisher's Exact test P=0.040), we sought to increase the detection of shared signatures of selection by considering overlapping windows in each of the BayeScan and BayPass analyses alone and taking the intersection. In each analysis, we extracted 10-kb genomic windows around significant SNPs (Bayes Factor (BF) > 30) and combined adjacent windows. 670 BayeScan 2 significant windows and 9195 BayPass significant windows overlapped between the Atlantic and Gulf clades. Intersecting these two window sets resulted in 259 'shared' candidate windows with significant signatures of directional selection and association with salinity in both clades (mean window size = 4593.2 bp). We used a randomization approach to generate a null distribution of the expected number of overlapping windows and bases given the size of the genome and windows. In 10,000 iterations, we shuffled the locations of significant windows in the genome and assessed the number of overlaps and intersecting bases using the Jaccard statistic in bedtools[4]. We found the empirical number of overlapping windows (N=259) and the Jaccard statistic (0.0233) were significantly greater than what was found in the randomized windows (Mean windows = 120.31 ± 0.204, P-value < 0.0001; Mean Jaccard = 0.00714 ± 1.47e-05, *P-value* < 0.0001).

II. Assessing the potential impact of mapping errors and bias

Sensitivity analyses were performed to assess whether the signatures of parallel directional selection and balancing selection could be explained by (1) mapping errors which could result from copy-number variants (CNVs) or difficult to align genomic regions, or (2) differences in mapping rates and accuracy between the Atlantic and Gulf Clades. As these analyses were computationally intensive, we focused these analyses only on the 'parallel' candidate SNPs.

a. Correlation between signatures of selection and mappability

GenMap[5] was used to calculate high-resolution mappability scores across the genome (read length 100, k=4). Mappability is a measure of the ability of a genomic region to produce a read that maps unambiguously to itself[6]. 'Parallel' candidate SNPs (Supplementary Box 1) were found to have high mappability scores with an average score of 0.997 vs the genome-wide average of 0.211 (One-tailed Wilcoxon, W = 2358700000, $P < 0.001$). The lowest mappability score of all candidate SNPs was 0.5. In a similar fashion, mappability scores for 10-kb windows around candidate SNPs did not differ from the genome-wide average of 10-kb window scores (Mean = 0.325, W=8923100, P=0.523). These results indicated that candidate SNPs were in highly unique and mappable genomic regions and that mapping errors were unlikely in these regions.

b. Correlation between signatures of selection and differential read depth

For our 'parallel' candidate SNPs there was no broad evidence for differential read coverage between the Atlantic and Gulf clade populations. For each population, read depth was calculated for every SNP using SAMtools v1.3 after filtering for low quality bases and alignments (Q < 20), removing duplicate reads with Picard v2.18.27 (http://broadinstitute.github.io/picard), and realigning around indels with GATK v3.8[7]. The support for differential coverage was estimated using the exact test in the R package *edgeR*[8]. The lowest *P*-value was 0.00291 and only 5 of the 349 candidate SNPs had a P-value < 0.05. None of these SNPs had a *P*-value < 0.05 after correcting for multiple tests using the Benjamini-Hochberg procedure[9]. There was no overall difference in *P*-values for candidate SNPs compared to all SNPs (Wilcoxon W= 2282400000, *P*-value = 0.682). The average $\log_2$ fold difference in coverage between the Gulf and Atlantic clade was only 0.155 for candidate SNPs, which was actually lower than the genome-wide SNP average of 1.68 (Wilcoxon W = 3382200000, *P*-value < 0.001).

A similar pattern was uncovered when analyzing coverage differences between the freshwater and saline populations (i.e. invasive and native populations). The mean $\log_2$ fold change for candidate SNPs was smaller than all SNPs (0.197 vs 0.211; Wilcoxon W = 3354300000, *P*-value < 0.001). There

was no difference in *P*-values for differential coverage compared to the rest of the genome (Wilcoxon W= 2276100000, *P*-value = 0.702). 23 SNPs did have a *P*-value < 0.05, but none were significant ($P < 0.05$) after correcting for multiple tests with the Benjamini-Hochberg procedure[9].

c. Correlation between signatures of selection and read depth

Genomic windows with aberrant coverage profiles could indicate the presence of CNV in Pool-Seq data[10]. While SNPs were only called at sites if all populations had a minimum and maximum coverage of 20 and 200, respectively (corresponding approximately to the 25% and 99% quantiles), it is still possible that genomic windows around candidate SNPs exhibit evidence of CNV that could mislead analyses. To investigate whether 10-kb windows around 'parallel' candidate SNPs exhibited evidence for CNV, we calculated read coverage for each position in the genome for each population using SAMtools v1.3, after filtering for low quality bases and alignments (Q < 20), removing duplicate reads with Picard, and realigning around indels with GATK.

For genomic windows (10 kb) around 'parallel' candidate SNPs, only one was an outlier in terms of mean coverage. We considered windows outside the 99% distribution of mean coverages of 10-kb windows outliers. The one outlier window on Scaffold 8 had extreme coverage in four of the nine populations and was removed from downstream analysis. This result indicated that the majority of candidate SNPs do not lie in genomic regions with considerably higher or lower sequencing coverage as might be expected from copy-number variants or difficult to map regions.

Together these results suggest that our candidate SNPs, signatures of parallelism, and associated signatures of balancing selection are not associated with mapping errors, differential mapping rates between the two clades, or copy-number variation. Given that the SNPs analyzed in our dataset were ones shared between the two clades with non-extreme sequencing coverage and high mapping quality, these SNPs appear to lie in regions of strong sequence conservation. Thus, while we do not suggest that copy-number variation is absent from our dataset, these analyses do suggest that CNV is unlikely to be driving the overall observed signals of parallelism and balancing selection at candidate SNPs. Future studies

should detect and quantify CNV between native and invasive population as potentially important variants underlying freshwater adaptation.

III. Signatures of balancing selection on shared variation vs private SNPs

Our finding that 'parallel' candidate SNPs were enriched for signatures of balancing selection in the native range prompts the question as to whether shared variation is broadly enriched for signatures of balancing selection. In other words, (1) Were all of the shared SNPs (N=366,781) enriched for signatures of balancing selection in the native range vs non-shared variation, and (2) could that explain why candidate parallel SNPs had elevated signatures of balancing selection?

We found that SNPs with a MAF > 0.05 in both clades (i.e. shared SNPs) indeed had stronger signatures of balancing selection on average than the rest of the genome across all four native saline populations (Montmagny: mean $\beta^{(2)} = 2.8$ vs 1.38, Wilcoxon $P < 0.001$; L'Isle Verte: mean $\beta^{(2)} = 2.25$ vs 0.828, Wilcoxon $P < 0.001$; Cocodrie Bayou: mean $\beta^{(2)} = 3.07$ vs 0.962, Wilcoxon $P < 0.001$; Taylor Bayou: mean $\beta^{(2)} = 3.02$ vs 0.963, Wilcoxon $P < 0.001$). This result was expected given that the $\beta^{(2)}$ score increases with a deficit of substitutions, and these SNPs were segregating in both clades. However, in all of these cases, $\beta^{(2)}$ scores for candidate 'parallel' SNPs were still significantly higher than those for all SNPs with a MAF > 0.05 (Montmagny: mean $\beta^{(2)} = 3.36$ vs 2.8, Wilcoxon $P < 0.001$; L'Isle Verte: mean $\beta^{(2)} = 2.84$ vs 2.25, Wilcoxon $P < 0.001$; Cocodrie Bayou: mean $\beta^{(2)} = 3.39$ vs 3.07, Wilcoxon $P < 0.001$; Taylor Bayou: mean $\beta^{(2)} = 3.34$ vs 3.02, Wilcoxon $P < 0.001$). These results suggested that all shared SNPs (likely representing ancient standing variation) do tend to harbor signatures of balancing selection, but the SNPs with signatures of parallel directional selection have especially elevated signatures of balancing selection even relative to all shared SNPs.

References Cited

1 Pickrell, J. K. & Pritchard, J. K. Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genetics* **8**, e1002967 (2012).

2 Foll, M. & Gaggiotti, O. A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: A Bayesian perspective. *Genetics* **180**, 977-993 (2008).

3 Gautier, M. Genome-wide scan for adaptive divergence and association with population-specific covariates. *Genetics* **201**, 1555-1579 (2015).

4 Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842 (2010).

5 Pockrandt, C., Alzamel, M., Iliopoulos, C. S. & Reinert, K. GenMap: Fast and exact computation of genome mappability. *bioRxiv*, doi:10.1101/611160 (2019).

6 Derrien, T. *et al.* Fast computation and applications of genome mappability. *PLoS One* **7**, e0030377 (2012).

7 McKenna, A. *et al.* The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* **20**, 1297-1303 (2010).

8 Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139-140 (2010).

9 Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate - a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B-Methodological* **57**, 289-300 (1995).

10 Schlotterer, C., Tobler, R., Kofler, R. & Nolte, V. Sequencing pools of individuals-mining genome-wide polymorphism data without big funding. *Nature Reviews Genetics* **15**, 749-763 (2014).