The genomic basis of adaptation in threespine stickleback fish

Inauguraldissertation

zur

Erlangung der Würde eines Doktors der Philosophie

vorgelegt der

Philosophisch-Naturwissenschaftlichen Fakultät

der Universität Basel

von

Quiterie Haenel

2022

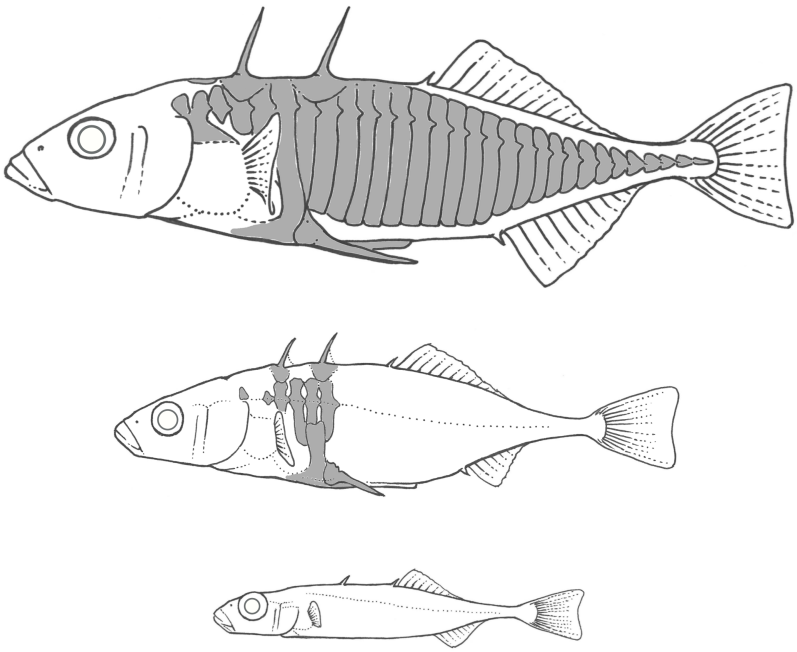
Genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultät auf Antrag von

Dr. Daniel Berner, Prof. Dr. Walter Salzburger, Prof. Dr. Yvonne Willi and Dr. Lukas Rüber

Basel, 15.09.2020

Prof. Dr. Martin Spiess
Dekan

The genomic basis of adaptation in threespine stickleback fish



Quiterie Haenel

Supervised by:

Dr. Daniel Berner, Prof. Dr. Walter Salzburger and Prof. Dr. Yvonne Willi

External expert:

Dr. Lukas Rüber

2022

Contents

1 Acknowledgements

I am extremely grateful to my supervisor Daniel Berner for trusting my motivation to do research in evolutionary biology no matter the background I had. His enthusiasm, patience, guidance and support helped me at each step of those last four years. I could not have imagined having a greater advisor and mentor for my PhD adventure. Thank you!

I am also grateful to Walter Salzburger for providing me the opportunity to join his lab and for his helpful support and advice, and to Lukas Rüber and Yvonne Willi for accepting to be a part of my PhD committee and for their enthusiasm.

Many thanks to past and present members of the Berner, Amrhein and Salzburger labs, with a special mention to Lilla Lovász and Nicolás Lichilín Ortiz for the stimulating discussions, support and fun. I also had a great pleasure working with all collaborators who contributed to this research work and in particular, Marius Roesti, Krista Oke, Andrew MacColl and Andrew Hendry.

I warmly thank the Freiwillige Akademische Gesellschaft (FAG) for their financial support that allowed me to complete my final PhD project and I would like to acknowledge Brigitte Aeschbach and Marianne Petrucci for their kind help when I was struggling with administrative issues...

Finally, I would like to thank my parents, my brother Médéric and Nolwenn, and my Mamie for their unconditional love and support, my friends that helped me disconnect from the PhD life during those four years, and Laurent, for everything, it was not always easy but you were always there for me... you are the best!

2 Introduction

Adaptation genomics

Evolutionary biology consists in the study of the evolutionary processes responsible of the diversification and adaptation of life forms over time. When adapting to a new environment (or changes in their local environment), populations have to adapt through natural selection. Until recently, the study of adaptation was focusing on fathoming the consequences of natural selection at the phenotypic level and how phenotypic evolution is linked to genetic changes. In the last 20 years, the development of new genetic and genomic tools, like high-throughput sequencing technologies, now allows the construction of reference genomes in a variety of non-model organisms and the investigation of the genomic basis of adaptation.

This was the aim of my PhD during those last four years and more specifically, I tried, within the main chapters of this thesis, to address the following questions exploring the consequences of natural selection at the molecular level:

- What is the genomic basis of parallel adaptation?
- Is adaptation polygenic? Does it happen mainly through the use of pre-existing or de novo mutations?
- How is standing genetic variation maintained?
- How can adaptation promote strong reproductive isolation?

Model species: the threespine stickleback

To answer those questions, one needs a suitable model organism. During my PhD, I used the threespine stickleback (Gasterosteus aculeatus), a small teleosts fish inhabiting the Northern hemisphere, that present many advantages when studying evolutionary biology. After the last glaciation retreat of the Pleistocene 10,000 to 12,000 years ago, marine stickleback colonized newly formed freshwater environments thus providing a wide range of aquatic habitats (i.e., small floodplain potholes, streams, lakes, estuaries) (Bell & Foster 1994). Local adaptation to those diverse aquatic habitats from standing genetic variation in marine stickleback led to numerous different phenotypes with variation of body form and external bony structures (i.e., dorsal and pelvic spines, lateral plates) (Bell & Foster 1994). Thus, present-day marine stickleback can also be considered as a surrogate for the ancestor as they tend to exhibit large effective population sizes and genetically well mixed over large distances (Mäkinen et al. 2006, Catchen et al. 2013, Roesti et al. 2014, Lescak et al. 2015, Galloway et al. 2020). Moreover, the availability of reliable reference genome (Jones et al. 2012), combined with a new online browser (https://stickleback.genetics.uga.edu/), and the fact that the genome is relatively small (460 Mb) make stickleback an excellent model to answer evolutionary biology question at the genomic level.

I studied predominantly two stickleback systems. First, stickleback from North Uist Island, Outer Hebrides, Scotland (UK) inhabiting multiple acidic and basic lakes and presenting striking ecotype differences (Waterston et al. 1979; Giles 1983; Spence et al. 2013; Klepaker et al. 2016; Magalhaes et al. 2016) for which we aimed to uncover the genomic basis of parallel adaptation and have some insights on how standing genetic variation is maintained in an ancestral population (see Main Chapters 1 and 2). The second system was the Misty watershed parapatric lake-stream pair of stickleback on Vancouver Island, British Columbia, CA, also presenting ecological and phenotypic differences between the lake and the stream (Lavin & McPhail 1993, Hendry et al. 2002) and our goal was to uncover how ecological divergence causes strong reproductive isolation in populations in close contact (see Main Chapter 3).

Approaches

Studying adaptation genomics is closely linked to the quality of the sequencing and the number of available markers. The more polymorphisms uncovered within the genome, the more efficient is the identification and quantification of regions or variants under adaptation. During my PhD, data were generated through two different techniques.

- Restriction-site Associated DNA (RAD) sequencing (Baird et al. 2008) where genomic DNA is cut with one or several restriction enzymes, randomly fragmented, then amplified, sequenced and aligned to a reference genome. With this method, only a part of the genome can be sequenced, dependent on the restriction enzymes used and the genome size. In this work (Main Chapter 1), we used an improved protocol based on two restriction enzymes (Nsi1 and Pst1) allowing the sequencing of approximately 1/3 of the stickleback genome.
- Whole-genome sequencing where genomic DNA is randomly fragmented, amplified, sequenced and aligned. With this method, we aimed to sequence the entire genome.

In both cases, it is important to consider an appropriate sample size, which combined with a high read depth will allow a precise estimation of allele frequencies (Ferretti et al. 2013, Gautier et al. 2013). The main analytical approach used to screen the genome of natural populations for signature of adaptation is the production of genome scans or "divergence mapping", assuming the availability of a reliable reference genome. These genome scans highlight the magnitude of differentiation between populations from different ecotypes and allow the identification of regions exhibiting a strong genomic differentiation and thus targeted by natural selection, also known as signatures of selection (Nielsen 2005, Storz 2005). In these particular regions, loci are under divergent selection (i.e., alternative alleles are selected in the different ecotypes). To measure divergence between populations, I used throughout my PhD the absolute allele frequency difference (AFD), a simple metric being a valuable alternative to FST (Wright 1950; Weir & Cockerham 1984; Holsinger & Weir 2009), as it exhibits a linear relationship from 0 to 1 along the allele frequency shift continuum (Berner et al. 2019). Main Chapters 1, 2 and 3 are using this approach.

On top of that, a whole-genome clinal analysis was performed (also used in Rafati et al. 2018) to study the genome-wide variation in a lake-stream stickleback habitat transition (Main Chapter 3). This consists in identifying genomic regions selected in an ecotype (here lake or stream) and looking at the allele frequency changes along the cline in these regions, likely implicated in reproductive isolation.

Thesis outline

During my PhD, I had the chance to collaborate with several people (see Acknowledgements and Chapters' headers) and to use powerful genomic approaches to study a combination of evolutionary biology topics that constitute the chapters of my thesis described below.

Chapter 1 (Haenel et al. 2019, Evolution Letters) explores parallel adaptation to acidic versus basic environment in North Uist lochs in the Outer Hebrides in Scotland. It investigates how ecological and phenotypic parallelism observed in those environments is reflected at the genomic level. Considering marine populations living around North Uist as a proxy for the marine ancestor, we demonstrated that basic-acidic differentiation occurred via the genome-wide sorting of standing genetic variation in the ancestor, with populations living in acidic (i.e. more extreme) derived habitats adapting through the accumulation of alleles rare in the ancestor and basic populations retaining alleles common in the ancestor.

Chapter 2 (Haenel et al. 2022, Molecular Ecology), closely linked to Chapter 1, investigates how standing genetic variation is maintained in ancestral populations. Two main theories exist: gene flow (genetic variants favored in novel habitats are disfavored in ancestral populations but maintained by continuous gene flow) and selective neutrality (genetic variants beneficial in novel habitats are essentially neutral in ancestral populations when they are rare). Based on the work described in Chapter 1, we considered five new marine samples across the Atlantic Ocean differing in geographic distance from North Uist and we explored the distribution and frequency of the acidic allele in the marine samples. We argued that when relatively rare, variants selected in derived habitats can persist selectively neutrally in the ancestor and not (only) because of gene flow between derived and ancestral populations.

Chapter 3 (Haenel et al. 2021, Nature Communications) aims to uncover, at the genomic level, how ecological divergence causes strong reproductive isolation between populations in close geographic contact. Considering a parapatric lake-stream stickleback in the Misty Lake watershed on Vancouver Island, BC, Canada, we performed a small-scale clinal analysis based on whole genome sequencing. We identified several regions fixed for alternative alleles as well as a steep cline in allele frequencies co-localizing with habitat transition thus suggesting that reproductive isolation is maintained by polygenic selection constituting a genome-wide barrier to gene flow without physical isolation.

Chapter 4 (Haenel, Laurentino et al. 2018, Molecular Ecology) consists in a review and is considered as a side chapter. It investigates the chromosome-scale distribution of crossovers, contributing to generate novel combination of alleles and thus have important evolutionary consequences, based on 62 animal, plant and fungal species. We highlighted that crossover rate in the center of chromosomes is strongly reduced compared to the peripheries, that heterogeneity in crossover rate is not systematically linked to centromere position and that the distribution of crossovers tends to be predicted by chromosome length. Moreover, those observations were consistent across the range of studied taxa. We thus argued about the importance of chromosome-scale heterogeneity in crossover rate into analytical tools in evolutionary genomics.

Four outreach chapters, describing the work initiated during my master internship and finalized during the PhD, conclude my thesis. This work focused on the use of metabarcoding techniques to uncover the biodiversity and community structure from different environments in complementarity with classical taxonomic approaches. In more details, Chapter 5 (Haenel et al. 2017) and Chapter 6 (Holovachov et al. 2017) concentrate on uncovering the diversity and community structure of the Swedish meiofauna based on sand and mud samples and comparing metabarcoding and classical taxonomic methods with a focus on marine nematodes. Chapter 7 (Jakubavičiūtė et al. 2017) compares classical taxonomic methods with metabarcoding to identify the diet of three-spined stickleback in the Baltic Sea. And finally, Chapter 8 (Ritter et al. 2019) investigates the richness pattern of birds, trees, eukaryotes and prokaryotes (insects, fungi and bacteria) across the Amazonian forests.

References

N. A. Baird et al. “Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers”. *Plos One* 3.10 (2008).

S. Bell M.; Foster. *The evolutionary biology of the threespine stickleback*. Oxford, UK: Oxford University Press, 1994, p. 571.

D. Berner. “Allele Frequency Difference AFD(-)An Intuitive Alternative to FST for Quantifying Genetic Population Differentiation”. *Genes (Basel)* 10.4 (2019).

J. Catchen et al. “The population structure and recent colonization history of Oregon threespine stickleback determined using restriction-site associated DNA-sequencing”. *Mol Ecol* 22.11 (2013), pp. 2864–83.

L. Ferretti, S. E. Ramos-Onsins, and M. Perez-Enciso. “Population genomics from pool sequencing”. *Mol Ecol* 22.22 (2013), pp. 5561–76.

J. Galloway, W. A. Cresko, and P. Ralph. “A Few Stickleback Suffice for the Transport of Alleles to New Lakes”. *G3-Genes Genomes Genetics* 10.2 (2020), pp. 505–514.

M. Gautier et al. “Estimation of population allele frequencies from next-generation sequencing data: pool-versus individual-based genotyping”. *Mol Ecol* 22.14 (2013), pp. 3766–79.

N. Giles. “The Possible Role of Environmental Calcium Levels during the Evolution of Phenotypic Diversity in Outer-Hebridean Populations of the 3-Spined Stickleback, Gasterosteus-Aculeatus”. *Journal of Zoology* 199.Apr (1983), pp. 535–544.

Q. Haenel et al. “Clinal genomic analysis reveals strong reproductive isolation across a steep habitat transition in stickleback fish”. *Nat Commun* 12.1 (2021), p. 4850.

Q. Haenel et al. “Meta-analysis of chromosome-scale crossover rate variation in eukaryotes and its significance to evolutionary genomics”. *Molecular Ecology* 27.11 (2018), pp. 2477–2497.

Q. Haenel et al. “NGS-based biodiversity and community structure analysis of meiofaunal eukaryotes in shell sand from Hallo island, Smogen, and soft mud from Gullmarn Fjord, Sweden”. *Biodiversity Data Journal* 5 (2017).

Q. Haenel et al. “Predictable genome-wide sorting of standing genetic variation during parallel adaptation to basic versus acidic environments in stickleback fish”. *Evolution Letters* 3.1 (2019), pp. 28–42.

Q. Haenel et al. “The maintenance of standing genetic variation: Gene flow vs. selective neutrality in Atlantic stickleback fish”. *Mol Ecol* 31 (2022), pp. 811–821.

A. P. Hendry, E. B. Taylor, and J. D. McPhail. “Adaptive divergence and the balance between selection and gene flow: lake and stream stickleback in the Misty system”. *Evolution* 56.6 (2002), pp. 1199–216.

O. Holovachov et al. “Taxonomy assignment approach determines the efficiency of identification of OTUs in marine nematodes”. *Royal Society Open Science* 4.8 (2017).

K. E. Holsinger and B. S. Weir. “Genetics in geographically structured populations: defining, estimating and interpreting F(ST)”. *Nat Rev Genet* 10.9 (2009), pp. 639–50.
E. Jakubaviciute et al. “DNA metabarcoding reveals diverse diet of the three-spined stickleback in a coastal ecosystem”. *Plos One* 12.10 (2017).
F. C. Jones et al. “The genomic basis of adaptive evolution in threespine sticklebacks”. *Nature* 484.7392 (2012), pp. 55–61.
T. Klepaker et al. “Selective agents in the adaptive radiation of Hebridean sticklebacks”. *Evolutionary Ecology Research* 17.2 (2016), pp. 243–262.
P.A. Lavin and J.D. McPhail. “Parapatric lake and stream sticklebacks on northern Vancouver Island: disjunct distribution or parallel evolution?” *Canadian Journal of Zoology* 71.1 (1993), pp. 11–17.
E. A. Lescak et al. “Evolution of stickleback in 50 years on earthquake-uplifted islands”. *Proc Natl Acad Sci U S A* 112.52 (2015), E7204–12.
I. S. Magalhaes et al. “The ecology of an adaptive radiation of three-spined stickleback from North Uist, Scotland”. *Molecular Ecology* 25.17 (2016), pp. 4319–4336.
H. S. Mäkinen, J. M. Cano, and J. Merilä. “Genetic relationships among marine and freshwater populations of the European three-spined stickleback (Gasterosteus aculeatus) revealed by microsatellites”. *Mol Ecol* 15.6 (2006), pp. 1519–34.
R. Nielsen. “Molecular signatures of natural selection”. *Annual Review of Genetics* 39 (2005), pp. 197–218.
N. Rafati et al. “A genomic map of clinal variation across the European rabbit hybrid zone”. *Mol Ecol* 27.6 (2018), pp. 1457–1478.
C. D. Ritter et al. “The pitfalls of biodiversity proxies: Differences in richness patterns of birds, trees and understudied diversity across Amazonia”. *Scientific Reports* 9 (2019).
M. Roesti et al. “The genomic signature of parallel adaptation from shared genetic variation”. *Molecular Ecology* 23.16 (2014), pp. 3944–3956.
R. Spence et al. “Ecological causes of morphological evolution in the three-spined stickleback”. *Ecology and Evolution* 3.6 (2013), pp. 1717–1726.
J. F. Storz. “Using genome scans of DNA polymorphism to infer adaptive population divergence”. *Molecular Ecology* 14.3 (2005), pp. 671–688.
A. Waterston et al. “The inland waters of the Outer Hebrides”. *Proceedings of the Royal Society of Edinburgh. Section B. Biological Sciences* 77 (1979), pp. 329–351.
B. S. Weir and C. C. Cockerham. “Estimating F-Statistics for the Analysis of Population-Structure”. *Evolution* 38.6 (1984), pp. 1358–1370.
S. Wright. “Genetical Structure of Populations”. *Nature* 166.4215 (1950), pp. 247–249.

3 Main Chapters

Chapter 1

Predictable genome-wide sorting of standing genetic variation during parallel adaptation to basic versus acidic environments in the stickleback fish

Haenel et al. 2019, Evolution Letters

LETTER

doi:10.1002/evl3.99

Predictable genome-wide sorting of standing genetic variation during parallel adaptation to basic versus acidic environments in stickleback fish

Quiterie Haenel,[1,2] Marius Roesti,[1,3,4] Dario Moser,[1,5] Andrew D. C. MacColl,[6] and Daniel Berner[1,7]

[1]*Department of Environmental Sciences, Zoology, University of Basel, 4051 Basel, Switzerland*

[2]*E-mail: quiterie.haenel@unibas.ch*

[3]*Biodiversity Research Centre and Zoology Department, University of British Columbia, Vancouver, British Columbia V6T 1Z4, Canada*

[4]*Current address: Institute of Ecology and Evolution, University of Bern, 3012 Bern, Switzerland*

[5]*Current address: Jagd- und Fischereiverwaltung Thurgau, 8510 Frauenfeld, Switzerland*

[6]*School of Life Sciences, University of Nottingham, Nottingham NG7 2RD, United Kingdom*

[7]*E-mail: daniel.berner@unibas.ch*

Received June 13, 2018
Accepted January 1, 2019

Genomic studies of parallel (or convergent) evolution often compare multiple populations diverged into two ecologically different habitats to search for loci repeatedly involved in adaptation. Because the shared ancestor of these populations is generally unavailable, the source of the alleles at adaptation loci, and the direction in which their frequencies were shifted during evolution, remain elusive. To shed light on these issues, we here use multiple populations of threespine stickleback fish adapted to two different types of derived freshwater habitats—basic and acidic lakes on the island of North Uist, Outer Hebrides, Scotland—and the present-day proxy of their marine ancestor. In a first step, we combine genome-wide pooled sequencing and targeted individual-level sequencing to demonstrate that ecological and phenotypic parallelism in basic-acidic divergence is reflected by genomic parallelism in dozens of genome regions. Exploiting data from the ancestor, we next show that the acidic populations, residing in ecologically more extreme derived habitats, have adapted by accumulating alleles rare in the ancestor, whereas the basic populations have retained alleles common in the ancestor. Genomic responses to selection are thus predictable from the ecological difference of each derived habitat type from the ancestral one. This asymmetric sorting of standing genetic variation at loci important to basic-acidic divergence has further resulted in more numerous selective sweeps in the acidic populations. Finally, our data suggest that the maintenance in marine fish of standing variation important to adaptive basic-acidic differentiation does not require extensive hybridization between the marine and freshwater populations. Overall, our study reveals striking genome-wide determinism in both the loci involved in parallel divergence, and in the direction in which alleles at these loci have been selected.

KEY WORDS: **Abiotic selection, convergence, ecological genomics, *Gasterosteus aculeatus*, North Uist, parallel evolution, Selective sweep, standing genetic variation.**

Impact Summary

The repeated emergence of similar life forms within ecologically similar environment provides particularly convincing evidence of determinism in evolutionary diversification driven by natural selection. While well documented at the phenotypic (i.e., trait) level, the genomic underpinnings of such parallel evolution remain elusive—to what extent is phenotypic parallelism reflected by genomic parallelism, and where do the genetic variants used for repeated adaptation originate? To examine these questions, we study young (postglacial) populations of stickleback fish displaying striking phenotypic similarity within multiple basic and acidic lakes on the island of North Uist, Scotland. We first type high-density genome-wide single-nucleotide polymorphisms (SNPs) five basic and five acidic populations, and in individuals from two marine sites, the latter providing a meaningful present-day surrogate of the genomic make-up of the marine ancestor of the lake populations. Based on these SNPs, we establish that the basic and acidic lake populations have adapted independently from one another. We then identify numerous genomic regions in which the populations show strong and consistent differentiation according to habitat, indicating widespread parallel genetic responses to divergent selection. Inspecting allele frequencies and allele associations in these regions reveals sharing of the same genetic variants within each habitat type. This adaptive genetic variation is also found in the marine ancestor, although variants selected in the ecologically relatively extreme acidic lakes tend to be uncommon in the sea. Nevertheless, these variants do not appear to be eliminated from the sea, likely because they are selectively (nearly) neutral when occurring at low frequency. Overall, our work highlights that phenotypic parallelism can be mirrored by parallel evolution at the genomic level; that the genome-wide sorting of standing genetic variation can be predicted from the ecological difference between novel and ancestral habitats; and that considering the ancestor can greatly strengthen genomic investigations of parallel evolution.

The quest for elucidating the genomic basis of adaptive diversification commonly proceeds by comparing populations from two ecologically distinct habitat types at genome-wide markers. Genetic loci important to differential adaptation are then identified by screening the populations for exceptionally strong habitat-related genetic differentiation relative to the genome-wide background level (e.g., Roesti et al. 2015; Lamichhaney et al. 2016; Reid et al. 2016; Yeaman et al. 2016; Marques et al. 2017). This approach is particularly informative when *multiple* populations adapted independently to each habitat type are available, as such "parallel" (or convergent; Arendt and Reznick 2008) evolution helps distinguish deterministic selective from stochastic genetic differentiation (Berner and Salzburger 2015).

Even deeper insights into the mechanisms of adaptation at the genomic level could, in principle, be gained by complementing the study of populations adapted in parallel to ecologically distinct habitats with genomic data from their shared ancestral population. The reason is that if information on the level of the ecological difference of each novel, derived habitat from the ancestral habitat is available, this allows generating a priori hypotheses about the direction and magnitude of selective genetic shifts away from the ancestor within each derived habitat—thus moving genomic analysis from description into the realm of prediction. Including the ancestor in genomic studies of multiple derived populations further offers the advantage that the origin of the polymorphisms under divergent selection between the derived habitats can be explored directly. An obvious obstacle to such genomic investigation, however, is that natural systems providing access to the ancestor of populations adapted in parallel to multiple derived habitats are rare.

We here adopt this uncommon analytical perspective in a genomic investigation of threespine stickleback fish from North Uist, an island of the Outer Hebrides, Scotland. Starting from a common marine ancestral population, stickleback have independently colonized numerous lakes on North Uist 8000–10,000 years ago (Fig. 1A; Campbell and Williamson 1979; Ballantyne 2010). Because of a sharp transition in surface geology, lakes in the west of the island display a basic pH and are meso- to eutrophic, whereas lakes in the east are consistently acidic, oligotrophic, and relatively depleted in dissolved ions (e.g., the calcium concentration is 10 times higher in the basic lakes than in the acidic lakes on average; Supporting Information Table S1). These ecological differences have proved stable over decades of investigation (Waterston et al. 1979; Giles 1983; Spence et al. 2013; Klepaker et al. 2016; Magalhaes et al. 2016). The difference in water chemistry between these two lake types, hereafter simply referred to as "basic" and "acidic," mirrors distinct levels of ecological difference from the ancestral marine habitat, with acidic lakes being more different from the sea than the basic lakes (visualized for pH and calcium concentration in Fig. 1B, top panels). Accordingly, basic and acidic lake populations have evolved different levels of phenotypic differentiation from their marine ancestor. For instance, marine stickleback generally exhibit long pelvic and dorsal spines and extensive lateral plating along most of their body, bony armor serving as protection from predators

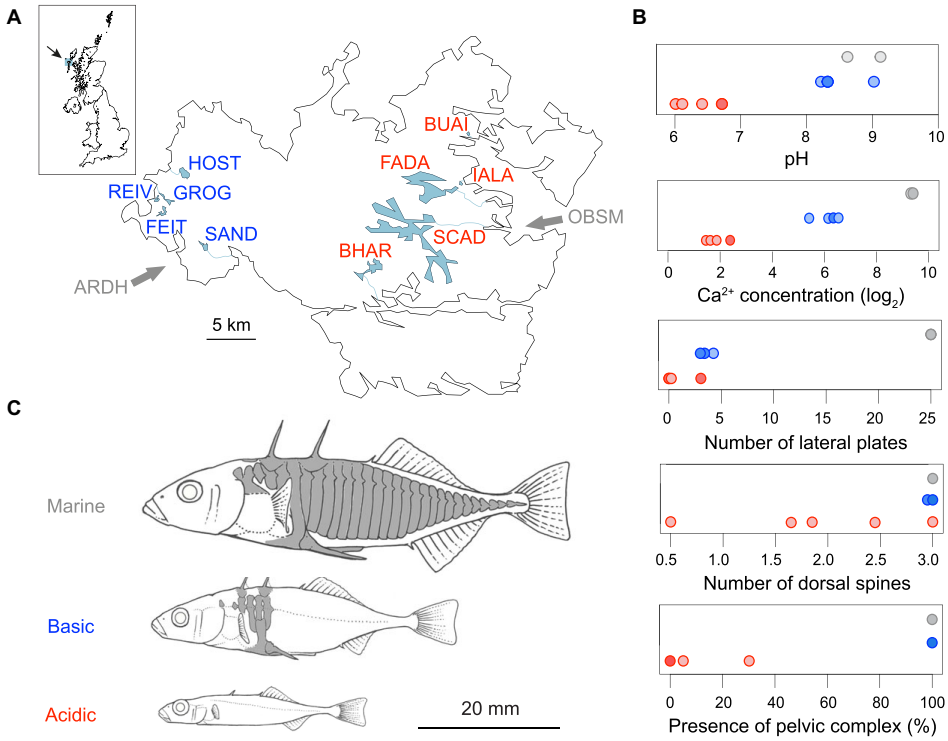


Figure 1. Stickleback study populations and their habitats. (A) Geographic situation of the basic (blue) and acidic (red) lakes on North Uist, Outer Hebrides, Scotland, with their connections to the sea shown as fine blue lines (the outlet of FEIT is uncertain). The gray arrows indicate the two coastal lagoons where marine stickleback were sampled. The same habitat-specific color coding is used throughout the paper. (B) pH, calcium (Ca^{2+}) concentration (in mg/L), and armor trait (lateral plate and dorsal spine counts, presence of pelvic spines and pelvic complex) mean values across individuals for each study site (data presented in detail in Supporting Information Table S1). The data points are arranged vertically according to habitat type, and they sometimes overlap (especially in the phenotypically uniform marine fish; overlap is indicated by darker dots). Data on pH and calcium concentration are from Magalhaes et al. (2016). Calcium measurements were $\log_2$ transformed. (C) Typical stickleback ecotypes from the three focal habitats, drawn to relative scale. Elements of bony armor are shaded in gray, including the dorsal spines, lateral plates, and the ventral pelvic complex to which the pelvic spines attach.

(Bell and Foster 1994). By contrast, typical freshwater ecotypes, including those in the basic lakes of North Uist, have their armor reduced to a bony girdle consisting of the pelvic complex and dorsal spines interconnected by only a few lateral plates (Fig. 1B and C; detailed data provided in Supporting Information Table S1). In stickleback ecotypes from the acidic lakes, this armor reduction has progressed further to an extreme level. Here, the pelvic complex, dorsal spines, and lateral plates are either rudimentary or missing altogether. Striking parallel evolution has also occurred in body size and shape (Fig. 1C; Campbell and Williamson 1979; Giles 1983; MacColl et al. 2013), with the dwarf stickleback residing in the acidic lakes ranking among the smallest vertebrates in Europe.

Based on the ecological differences between the sea, the basic lakes and the acidic lakes, and the concurrent phenotypic parallelism exhibited among the derived populations within each freshwater habitat, we ask two main questions guiding our genomic investigation: first, is parallelism in the evolution of basic and acidic lake stickleback ecotypes mirrored in the sharing of distinctive adaptive genetic variants within each lake habitat type? Despite its simplicity and importance to understanding the genomic basis of evolution, the congruence in parallelism at the phenotypic and genotypic levels remains poorly evaluated empirically in higher organisms. The reason is that genomic investigations of natural systems exhibiting phenotypic parallelism commonly lack the marker resolution needed to achieve robust

conclusions about parallelism at the genomic level (Berner and Roesti 2017; Lowry et al. 2017; Haenel et al. 2018; for exceptions, see Martin et al. 2013; Lamichhaney et al. 2016; Reid et al. 2016; Yeaman et al. 2016; Elgvin et al. 2017). Second, has the greater ecological difference of the acidic than basic lakes from the ancestral habitat caused asymmetry in the selection of adaptive genetic variation? We expect that stickleback in acidic lakes should have adapted to their extreme habitat by accumulating alleles relatively rare in their ancestor, while at the same loci, the populations in basic lakes should have retained alleles also occurring at high frequency in the sea. This idea is amenable to empirical examination because present-day marine stickleback living around North Uist provide a proxy for the ancestor of all derived freshwater populations on the island. Using high-resolution single-nucleotide polymorphism (SNP) data from samples from all three habitats, we confirm both genomic parallelism and habitat-related asymmetry in the selection of standing genetic variation, thus uncovering a strongly deterministic component to adaptive diversification at the genomic level.

Methods

STICKLEBACK SAMPLES

Freshwater stickleback were captured from five basic and five acidic lakes on North Uist (Fig. 1A) during the 2014 breeding season (mid-April to late May), aiming for a sample size of 30 individuals per lake (details given in Supporting Information Table S1). The lakes were chosen to represent separate watersheds draining independently into the sea, although for one basic lake, this could not be determined unambiguously (FEIT; this lake may reside in the same watershed as GROG, although a direct present-day connection can be ruled out). Marine stickleback were sampled on breeding grounds in two tidal lagoons located on the east coast (OBSM, $N = 20$, sampled 2013) and west coast (ARDH, $N = 10$, sampled 2016) of the island. These fish, however, were not lagoon-residents—which exist on North Uist but are phenotypically distinct (El Nagar and MacColl 2016)—but truly anadromous marine stickleback. Marine stickleback exhibit large population sizes and are genetically well mixed over large distances (Hohenlohe et al. 2010; Jones et al. 2012a; Catchen et al. 2013; Roesti et al. 2014; Lescak et al. 2015), and they inhabit a relatively constant habitat, so that present-day marine samples are generally considered meaningful surrogates of the ancestor of nearby derived freshwater populations. All sampling was performed with unbaited minnow traps. Specimens were euthanized with an overdose of MS222 and immediately preserved in absolute ethanol. Details on the sampling locations and habitats are given in Supporting Information Table S1 (see also Giles 1983; MacColl et al. 2013; Spence et al. 2013; Klepaker et al. 2016; Magalhaes et al. 2016).

PHENOTYPIC ANALYSES

To highlight parallelism in phenotypic evolution among our study populations from each habitat type, we scored armor traits known to exhibit strong variation among North Uist freshwater stickleback, presumably driven by selection associated with predation and differences in water chemistry (Giles 1983; MacColl et al. 2013; Spence et al. 2013; Klepaker et al. 2016; Magalhaes et al. 2016). These traits were chosen for ease of measurement, recognizing that selection related to water chemistry has likely targeted numerous life history and physiological traits beyond external bone morphology. Twenty total individuals chosen at random from each lake sample, and all individuals from the two (smaller) marine samples, were scored under a dissecting microscope for the number of lateral plates (right body side), number of dorsal spines, presence (at least as rudiment) or absence of the pelvic complex and of the pelvic spines. Counts were averaged for each population.

DNA LIBRARY PREPARATION AND SEQUENCING

To obtain genetic markers, we first extracted DNA individually from fin tissue of each of the 288 total stickleback from the 10 total freshwater populations (Supporting Information Table S1) by using a MagNA Pure LC278 extraction robot (Roche, Basel, Switzerland) and the Tissue Isolation Kit II. After an RNase treatment, DNA concentrations were standardized to 10 ng/μL based on two rounds of Qubit (Invitrogen, Thermo Fisher Scientific, Wilmington, DE, USA) quantitation, and used to prepare high-resolution pooled restriction site-associated DNA (RAD). Specifically, 3.3 μL of adjusted DNA solution from each individual of a given population were transferred to each of two replicate sample pools. Each of these two pools per population was then split further into two subpools of 50 μL subjected to restriction with either the *Nsi1* enzyme (approximately 164,000 recognition sites across the 460 megabases [Mb] stickleback genome) or the *Pst1* enzyme (314,000 recognition sites) (New England Biolabs, Ipswich, MA, USA). The rationale of the parallel restriction of each subpool with a separate enzyme was to avoid DNA fragments too short for sequencing that would have resulted from the *simultaneous* restriction with *Nsi1* and *Pst1* at recognition sites located in close proximity. Our parallel-restriction approach (see Supporting Information Fig. S1 for a schematic) thus allowed interrogating the stickleback genome at approximately 478,000 total restriction sites, resulting in a 22 times higher physical resolution than what would be achieved by using the standard *Sbf1* enzyme. The two digested subpools of each pool were then labeled with the same molecular barcode (four barcodes used in total; two 5mer, one 6mer, and one 7mer) and then combined, yielding two replicate

pools per population. These pools were then subjected to the standard RAD library preparation steps (Baird et al. 2008). Enrichment polymerase chain reaction (PCR) occurred in seven replicate reactions per library (i.e., pool) to reduce amplification bias. The 20 total libraries were single-end sequenced to 200 base pairs on five lanes of an Illumina HiSeq2000 instrument, always allocating the two replicate libraries of a given population to different lanes.

DNA from the 30 total individuals from the two marine samples was extracted (Qiagen DNeasy Blood & Tissue Kit Valencia, CA, USA) and barcoded individually, pooled PCR-free into a single library, and whole-genome (not RAD) paired-end sequenced to 151 base pairs on a single Illumina HiSeq2500 lane.

MARKER GENERATION

Raw sequence reads were parsed by barcode (i.e., population), pooled over the two replicate libraries, and aligned to the third-generation assembly (Glazer et al. 2015) of the stickleback reference genome (Jones et al. 2012b) by using Novoalign (Version 3.0, http://www.novocraft.com/products/novoalign/; alignment settings provided on the Dryad repository, https://doi.org/10.5061/dryad.4ck2q0m). Resulting SAM files were converted to BAM format and accessed in R using Rsamtools (Morgan et al. 2017). SNPs were ascertained in the global freshwater pool (i.e., all basic and acidic populations combined), requiring a total read coverage between 150 and 2800 (the latter effectively filtering sequences from repeated elements), a minor allele frequency (MAF) superior to 0.05 across the pool, and a distance to the nearest polymorphism of at least 12 base pairs (effectively avoiding microindel stutter). A total of 253,451 SNPs passed these filters, yielding an approximate average resolution of 1 SNP per 2 kilobases (kb)—higher than in any previous reduced-representation sequencing study (Lowry et al. 2017). At these SNPs, we performed nucleotide counts for each freshwater population at an average read depth of 63× per population pool. At the same SNPs, we then also performed nucleotide counts for the two marine samples (for which full-genome data were generated; Supporting Information Fig. S1). For analysis, nucleotide counts from all samples were stored in a single SNP matrix (available on Dryad). To achieve the standard individual-level sample size of the lake populations, SNP data from the two genetically very similar (Supporting Information Tables S2 and S3) marine samples were combined to a single population (average read depth: 133×) in all analyses except the phylogenies and ordination (a detailed justification for combining the two marine samples to a single population is provided in the Supporting Information "Discussion" section, paragraph 1).

GENETIC SIMILARITY AMONG POPULATIONS

A fundamental requirement in investigations of parallel evolution is evidence that the focal populations have adapted to their habitat independently from other, ecologically similar populations (Endler 1986; Schluter 2000). Our study lakes presently reside in separate watersheds draining independently into the sea (Fig. 1A). Moreover, a previous genetic investigation including a subset of our study populations suggests their evolutionary independence (Magalhaes et al. 2016). To extend this evidence to all our study populations, we first characterized their genetic similarity by nuclear phylogenies. For this, we filtered the SNP matrix for SNPs occurring alone on a RAD locus (i.e., "loner" SNPs sensu Roesti et al. 2015; to maximize their independence) and exhibiting a base coverage of at least 40× within each population pool. To minimize the influence of selection, we further excluded SNPs showing substantial allele frequency differentiation (>0.5) in both the combined basic-acidic and marine-freshwater differentiation scans (details below). Moreover, a SNP had to reside within 5 Mb from the nearest tip of the corresponding chromosome—a genomic region exhibiting a particularly high recombination rate (Roesti et al. 2013; Glazer et al. 2015). Since we sequenced pooled DNA and hence individual genotypes were not available, we used the resulting 15,058 SNPs to generate 10 synthetic diploid genotypes for each population by drawing nucleotides at random without replacement from the corresponding population pool and concatenating them after translation to IUPAC ambiguity code. We next inferred the most appropriate model of sequence evolution ("GTR+G+I") using the R package *phangorn* (Schliep 2011) and constructed a maximum likelihood tree (neighbor-joining produced similar results in all phylogenetic analyses). In this analysis, all 10 freshwater populations proved reciprocally monophyletic—consistent with the absence of admixture inferred from individual-level genotype data from a subset of our populations including all acidic ones (Fig. 3 and Supporting Information Fig. S1 in Magalhaes et al. 2016), so we present a simplified tree based on a single individual per population only (data provided in fasta format on Dryad; the tree based on the full samples is shown in Supporting Information Fig. S2). Additional phylogenies were performed by expanding the dataset to *all* loner SNPs satisfying the above base coverage criterion (68,245 SNPs), and by restricting the dataset to loner SNPs separated by at least 1 Mb (227 SNPs). Over this latter physical distance, linkage disequilibrium is minimal in this (e.g., Roesti et al. 2015) and many other species (Lowry et al. 2017) so that synthetic multilocus genotypes should resemble natural genotypes (further justification for using synthetic genotypes for phylogenetic inference is elaborated in the Supporting Information "Discussion" section, paragraph 2).

In a second analysis, we explored the genetic similarity among our populations by ordination using nonmetric multidimensional scaling (NMDS) and the stringently filtered dataset described above (15,058 SNPs). At each SNP, we first identified the major and minor allele across the global allele pool comprising all freshwater populations. Next, we randomly sampled a single allele from each population at each SNP, assigned these alleles

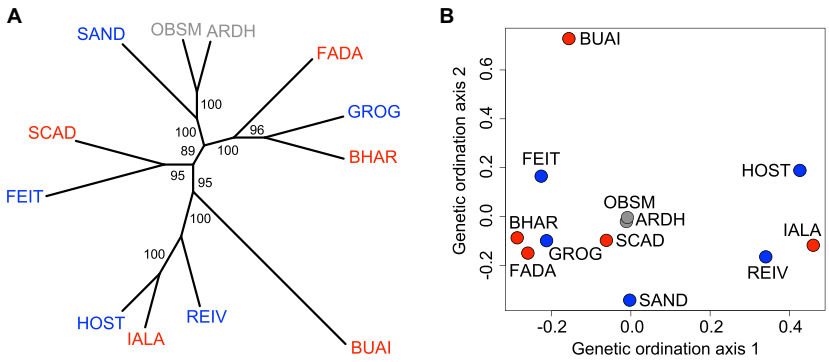


Figure 2. **(A) Unrooted nuclear maximum likelihood tree based on 15,058 loner SNPs located in the high-recombination chromosome peripheries and showing low AFD in between-habitat population comparisons, using a single synthetic stickleback individual per population. Color coding is by habitat, as in Figure 1. The numbers give bootstrap support for all nodes. Additional trees based on multiple synthetic individuals per population, neighbor joining, the full genome-wide set of loner SNPs, or just loner SNPs spaced by at least 1 Mb are shown in Supporting Information Figures S2, S9, S10, and S11). (B) Genetic similarity among the populations shown by their position along the first two NMDS ordination axes.**

the value of 1 (major allele) or 0 (minor allele), and derived a binary population similarity matrix from these values using the R function *dist*. Finally, we extracted ordination coordinates from the similarity matrix using the function *isoMDS* (good fit was achieved with seven dimensions; stress = 0.06), and visualized the populations along the first two. Running a principal component analysis on the same data and visualizing the populations along the first two components produced almost identical results.

IDENTIFYING LOCI UNDER PARALLEL BASIC-ACIDIC DIFFERENTIATION

A major objective of our study was to assess if the repeated evolution of characteristic ecotypes within the basic and acidic lakes is mirrored by the consistent, parallel sorting of genetic variation between these habitats. Our key resource to address this question were genome-wide scans for the magnitude of genetic differentiation performed for all 45 possible pairwise population comparisons within and across the freshwater habitat types. These included 25 basic-acidic (B-A), 10 basic-basic (B-B), and 10 acidic-acidic (A-A) population combinations. We here considered only SNPs displaying a total read coverage of at least 50× within each population, and a MAF across the global pool of all B and A populations superior to 0.2 to ensure adequate information content (Roesti et al. 2012). Although not the focus of the present study, we also performed an analogous genome-wide differentiation scan by treating all lake populations simply as freshwater stickleback, and comparing them to our marine population (i.e., a standard analysis of parallel evolution in marine-freshwater stickleback; e.g., Hohenlohe et al. 2010; Jones et al. 2012a; Roesti et al. 2014; Terekhanova et al. 2014). As a resource, this latter analysis is presented as Supporting Information Figure S3 and the underlying SNP data (ascertained differently than our main SNP dataset) are provided on Dryad.

Population differentiation, quantified by the absolute allele frequency difference (AFD), was then integrated within each of the three freshwater habitat comparison categories. To do so, we calculated at each SNP the mean of the AFD values from all population pairs in a given comparison category, provided the SNP was represented by a sufficient number of individual comparisons (thresholds: at least 18 for B-A, at least 8 for B-B and A-A). (Genome-wide mean and median differentiation values for all pairwise population and habitat type comparisons, quantified by both AFD and F_{ST}, are presented in Supporting Information Tables S2 and S3, and the underlying raw pairwise comparisons are available on Dryad). This averaging did not involve adjusting AFD values from a given population comparison by the corresponding overall level of differentiation, although performing such standardization did not materially affect our results (Supporting Information Fig. S4). Integrated this way, the AFD data were screened for genomic regions displaying exceptionally strong and consistent differentiation in the B-A comparison category (note that this approach necessarily precludes conclusions about genomic regions involved *inconsistently* in adaptation within an ecotype; see Discussion S2 in Roesti et al. 2014). We identified genomic regions of extreme B-A differentiation based on all SNPs exceeding an AFD threshold of 0.70, corresponding to the 99.95 percentile of the AFD distribution across all genome-wide SNPs in this comparison category (204,433 SNPs). When located on the same chromosome, a high-differentiation SNP was considered to represent an independent genome region when

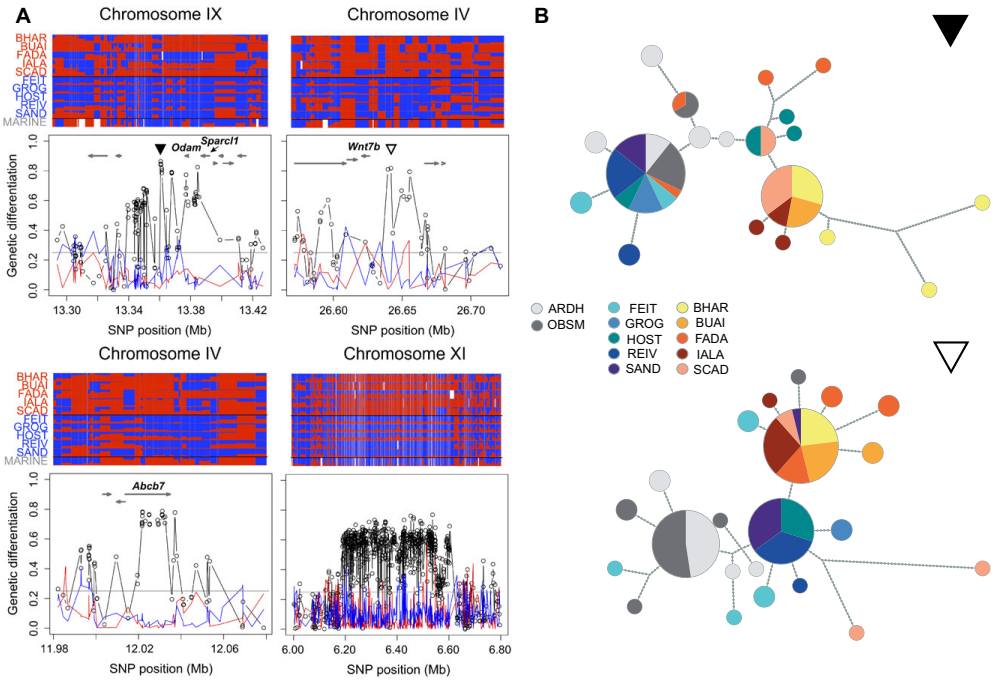


Figure 3. **(A) Four exemplary genomic regions around core SNPs showing strong and highly parallel differentiation between basic and acidic stickleback ecotypes. The bottom panels show mean genetic differentiation (absolute allele frequency difference, AFD) profiles for the integrated B-A (black), B-B (blue), and A-A (red) population comparisons. The dots represent individual SNPs, and the horizontal gray lines indicate genome-wide median AFD for the integrated B-A comparisons. Gray arrows show the location of genes (not for the large inversion on chromosome XI), with four candidates for B-A adaptation labeled. The top panels summarize allele frequencies for each population at all SNPs underlying the AFD profiles on the bottom. Alleles are color coded in blue (basic alleles predominant in the basic ecotype pool) and red (acidic alleles). Cell widths are delimited by the midpoints between each focal SNP and its flanking SNPs. White cells represent missing data. (B) Haplotype genealogies based on SNPs from targeted individual-level sequencing at the two top core SNPs. The position of the target segments is indicated by a filled (upper genealogy) and empty (lower genealogy) black triangle in (A). Each pie represents a unique haplotype (or a collection of closely related haplotypes, as these were collapsed; see "Methods" section), and edges connecting pies or nodes indicate one inferred mutational step.**

separated by at least 50 kb from any other such SNP. We hereafter refer to the single most strongly differentiated SNP within each high-differentiation region thus identified as "core SNP." In a supplementary analysis, the exactly same SNPs underlying our integrated B-A comparison were subjected to a search for habitat-associated outliers using BayPass (Gautier 2015), which generally identified similar genomic regions as our method (Supporting Information Fig. S5A). We next retrieved all genes located within a 100 kb window centered at each core SNP from the reference genome annotation, along with their functions as specified by the Ensembl and GeneCards data bases. This information was not subjected to a formal candidate gene analysis, but inspected qualitatively for genes appearing particularly likely to be involved in bone evolution, or having emerged as candidate adaptation genes in previous stickleback work.

To support the reliability of our search for genomic regions involved in parallel B-A differentiation based on pooled RAD sequencing and the averaging of multiple population comparisons, we performed targeted individual-level Sanger sequencing at two top core SNPs identified by the above genome scans. For both regions, we amplified a 700 bp fragment from a subsample of 4–8 individuals per sample site (Supporting Information Table S1), using primers and PCR conditions described in the Supporting Information "Methods". The resulting sequences were aligned and screened for SNPs using Geneious version 11.1.2, and haplotype reconstruction was performed using PHASE version 2.1

(Stephens et al. 2001). Genealogies were then constructed with RAxML version 8 (Stamatakis 2014) and visualized as haplotype networks in FITCHI (Matschiner 2016) by collapsing haplotypes separated by less than three edges (–e 3 option).

CHARACTERIZING THE ALLELES UNDER PARALLEL BASIC-ACIDIC DIFFERENTIATION

Acidic lakes show a greater ecological difference from the sea than basic lakes, and acidic ecotypes display stronger phenotypic differentiation from their marine ancestor than basic ecotypes (Fig. 1B and C). Our second main expectation was thus that at loci showing parallel B-A differentiation, the acidic ecotypes should generally have recruited alleles relatively unfavorable and hence rare in the marine ancestor, whereas the basic ecotypes should have retained alleles common in the ancestor. This prediction was investigated by three approaches based on the core SNPs identified as described above ($N = 42$). The first approach was phylogenetic and involved deriving a single diploid multilocus genotype for each population by sampling two nucleotides at random from each population-specific pool at each core SNP, and concatenating them as IUPAC characters. The resulting data were used to construct a maximum likelihood tree as described above. We then repeated this procedure for the same number of SNPs ($N = 42$) chosen at random from the genome-wide SNP panel. To ensure that these latter "random SNPs" were minimally affected by divergent selection between the basic and acidic lakes, we here only considered SNPs exhibiting a magnitude of differentiation within 0.5% of the median value (AFD = 0.25) observed across all SNPs in the combined B-A comparison. Our prediction in this phylogenetic analysis was that at the core SNPs, the basic populations should show a greater genetic similarity to the marine fish than the acidic populations, whereas in the genealogy for the random SNPs, no freshwater ecotype should appear systematically closer to the marine fish.

Our second approach to investigating if the acidic ecotypes are genetically more derived from their marine ancestor than the basic ones involved ordination using NMDS. We here followed the protocol described above, except that only a single ordination axis was extracted from both the core and random SNPs. Our prediction was that at the core SNPs only, the basic populations should display a greater genetic similarity, and hence greater proximity on the NMDS ordinate, to the marine fish than the acidic populations.

The third approach, finally, was a locus-specific analysis of allele frequencies at the core SNPs. We here first classified the two alleles at each SNP as "basic" or "acidic," based on their average frequency over all populations within each lake type. That is, the allele exhibiting an average frequency >0.5 across the basic populations was considered the basic allele, and vice versa. Then we determined the frequency of the basic and acidic allele at each SNP in the marine population, which allowed us to evaluate the prediction that core SNP alleles characteristic of the acidic ecotypes occur at relatively low frequency in the marine ancestor. We here again used the random SNPs as a negative control, determining basic and acidic alleles as described for the core SNPs.

As a robustness check, all analyses described in this section were repeated with an independent sample of random SNPs selected by controlling their magnitude of differentiation in the B-A comparison less strictly. This produced very similar results supporting the same conclusions (Supporting Information Fig. S6).

ANALYSIS OF SELECTIVE SWEEPS

Observing that core SNP alleles typical of acidic ecotypes tended to be less common than basic alleles in the ancestral habitat (see "Results and Discussion" section), we finally hypothesized that genetic diversity should be relatively reduced around the core SNPs in the acidic populations. The reason is that in these populations, the locally favorable variants must generally have experienced greater frequency changes reducing neutral variation in the physically linked chromosomal neighborhood particularly effectively (i.e., stronger selective sweeps) (Maynard Smith and Haigh et al.1974). To explore this idea, we quantified genetic diversity within each lake population as the total number of SNPs with a MAF >0.3 across the 40 kb window surrounding a given core SNP. A high MAF threshold was chosen because selective sweeps shift the MAF distribution downward (Braverman et al. 1995), hence the density of high-MAF SNPs should be particularly sensitive to sweeps. A supplementary analysis comparing the density of high-MAF SNPs to nucleotide diversity (π) as measures of genetic diversity confirmed this expectation, and further revealed that the former is highly robust to prefiltering SNP data with mild MAF thresholds whereas nucleotide diversity can become strongly biased by such filtering (Supporting Information Fig. S7). The SNP count obtained was then summed over all populations within each lake category and divided by the analogous sum of SNPs observed across a larger (1 Mb) window around the same core SNP. The latter standardization served to adjust for general differences in genetic diversity between the ecotypes. For the relative "SNP density" metric thus obtained for each core SNP, we next calculated the difference between the basic and the acidic habitat. Finally, we evaluated if this B-A difference in SNP density was related to the frequency of the acidic allele in the marine population. Our expectation was a negative relationship, indicating a particularly strong reduction in genetic diversity in the acidic populations at those core SNPs at which the acidic allele had to rise from particularly low initial frequency during adaptation. The random SNPs were again used analogously as a

negative control. A robustness check for this analysis of selective sweeps is presented in Supporting Information Figure S8.

Unless specified otherwise, all analyses were performed with the R language (R Core Team 2017; codes for the main analyses are available on Dryad). Variation around estimated statistics was quantified through bootstrapping (Manly 2006) with 10,000 iterations.

Results and Discussion

BASIC AND ACIDIC STICKLEBACK ECOTYPES ON NORTH UIST HAVE EVOLVED INDEPENDENTLY

In our nuclear SNP phylogeny based on synthetic genotypes, basic and acidic populations appeared well-mixed across the genealogical tree. Some terminal bifurcations, for instance, involved a basic and an acidic population originating from geographically distant lakes (HOST-IALA, BHAR-GROG) (Fig. 2A and Supporting Information Figs. S2, S9, S10, and S11). Conversely, populations from lakes located in closest geographic proximity and belonging to the same ecotype (GROG-REIV and FADA-IALA) appeared on distinct basal branches of the tree. Ordination of the populations also indicated the absence of genetic similarity by ecotype (Fig. 2B). (See also the weak correlations in allele frequencies estimated by BayPass for all population combinations except FEIT and GROG, Supporting Information Fig. S5B; these two populations may not qualify as fully independent.) These genetic patterns, combined with the geographic separation of the basic and acidic habitats due to surface geology, render the major alternative to parallel evolution—the single origin of a basic and an acidic ecotype followed by admixture between the ecotypes during secondary contact in different localities (Bierne et al. 2013)—highly implausible. Instead, our analyses support the view that our freshwater populations were founded independently by ancestral marine stickleback and then evolved in isolation from each other, consistent with the present-day hydrological independence of the lakes (Fig. 1A) (further support for the conclusion of the evolutionary independence of our freshwater populations is elaborated in the Supporting Information "Discussion" section, paragraph 3). The multiple phenotypically similar populations within each lake type are thus well suited for an investigation of the genomics of parallel adaptation of an ancestral population to two ecologically distinct derived habitats.

BASIC AND ACIDIC POPULATIONS HAVE DIVERGED IN PARALLEL IN NUMEROUS GENOMIC REGIONS

Having obtained strong evidence that our focal freshwater stickleback populations adapted in parallel to basic and acidic lakes, our first main objective was to search for genomic regions playing a key role in the differentiation between basic and acidic ecotypes. After combining AFD data from all possible comparisons between basic and acidic lakes (B-A comparisons), 42 independent genomic regions satisfied our criteria for loci under highly parallel B-A differentiation. These regions generally contained multiple SNPs nearly fixed for alternative alleles between most populations from the two lake types (four examples are presented in Fig. 3A; the top [core SNP showing AFD > 0.75] 19 regions are visualized in detail in Supporting Information Fig. S12, and a genome-wide differentiation plot is presented in Supporting Information Fig. S13). In these regions, we generally observed low differentiation *within* each habitat type (i.e., in the B-B and A-A comparisons; Fig. 3A and Supporting Information Fig. S12), indicating extensive haplotype sharing within each ecotype and hence ruling out the possibility that the populations adapted by selecting independent new mutations in these genomic regions (Roesti et al. 2014; Berner and Salzburger 2015). These conclusions—derived from pooled sequencing data—were supported by our individual-level targeted sequencing at two top core SNPs: in both genomic regions, we observed that basic and acidic ecotypes formed distinct haplotype clusters, and that haplotypes identical by descent were shared among multiple populations within each ecotype (Fig. 3B).

A qualitative inspection of the 100 kb windows around the core SNPs suggested potential candidate genes for adaptive differentiation between basic and acidic environments. These included *Sparcl1* and *Odam* for the core SNP region on chromosome IX (Fig. 3A), both involved in vertebrate tissue mineralization and specifically bone and tooth development (Kawasaki et al. 2004; Kawasaki 2009). Other suggestive candidates were *Wnt7b* and *Abcb7* on chromosome IV (Fig. 3A). These latter genes have been suggested to be under divergent selection between marine and freshwater stickleback (Jones et al. 2012a; Jones et al. 2012b; Roesti et al. 2014; see also Supporting Information Fig. S3A), but here also appear involved in the differentiation between ecologically different *freshwater* habitats. A complete list of genes around the top core SNPs is presented as Supporting Information Table S4.

The core SNP regions also included an inversion of several hundred kilobases on chromosome XI (Fig. 3A), a locus commonly found highly differentiated between marine and freshwater stickleback (Hohenlohe et al. 2010; Jones et al. 2012b; Roesti et al. 2014; Supporting Information Fig. S3), but also between stickleback residing in adjoining lake and stream habitats exhibiting very similar water chemistry (Roesti et al. 2015). Divergent selection on this inversion across qualitatively different habitat transitions poses a major challenge to understanding what loci captured by the inversion are actually fitness-relevant in each ecological context. The B-A comparisons revealed two further genomic regions (on chromosomes V and XVII) showing extended population differentiation over hundreds of kilobases, although the consistency of differentiation across population

comparisons was lower (Supporting Information Fig. S14). These regions further exhibited distinct MAF strata consistent with inversions (Roesti et al. 2015), but not the distortion in read alignment success characteristic of *ancient* inversion polymorphisms showing massive sequence differentiation (Roesti et al. 2013) (Supporting Information Fig. S14). We thus speculate that these regions may be relatively young inversions.

BASIC-ACIDIC DIFFERENTIATION THROUGH ASYMMETRIC SELECTION OF STANDING GENETIC VARIATION

The identification of genomic regions selected differentially between the two types of derived freshwater habitats motivated our second main prediction: that at loci of strong B-A differentiation, the acidic ecotypes, residing in habitats more ecologically different from the ancestral marine habitat than the basic ecotypes, tend to have accumulated alleles uncommon in the ancestor. This prediction was confirmed by our phylogeny based on the core SNPs representing the 42 genomic regions of high B-A differentiation (these core SNPs are characterized in detail in Supporting Information Table S5): in the genealogical tree, the basic populations formed a distinct branch clustering closely with the marine samples, whereas all acidic populations together formed a separate branch highly distinct from the one including the marine and basic fish (Fig. 4A left and Supporting Information Fig. S6A; see also Fig. 3B top for similar evidence based on individual-level haplotype data). Consistent with the genome-wide nuclear phylogeny (Fig. 2A), however, the genealogy based on the 42 random SNPs did not indicate a stronger genetic similarity of the basic than acidic ecotypes to the marine ancestor (Fig. 4A right). Likewise, our ordination analysis using the core SNPs revealed a close genetic similarity between basic and marine stickleback, with the acidic ecotypes appearing very different from these two (Fig. 4B top and Supporting Information Fig. S6B). By contrast, ordination based on the random SNPs indicated no genetic structure by habitat (Fig. 4B bottom).

These insights were refined by the inspection of allele frequencies within each population. Classifying the two alleles at the core SNPs as basic or acidic based on their overall frequency within the two freshwater habitats, we first observed that at these SNPs, the marine ancestor consistently harbored *both* alleles (Supporting Information Table S5). However, the acidic allele was the less common (i.e., the *minor* allele, frequency <0.5) in the marine fish at 32 of the 42 SNPs (two-tailed binomial probability of an asymmetry of this magnitude or greater: $P = 0.0009$). The frequency distribution of the acidic alleles across the core SNPs was thus biased toward low values in marine stickleback and differed strikingly from uniformity expected for polymorphisms having segregated under selective neutrality for a long time (Wright 1931) (Fig. 4C left and Supporting Information Fig. S6C). The latter frequency distribution, however, was observed for the acidic alleles at the random SNPs (Fig. 4C right; the frequency distributions of the core and random SNPs differed clearly: $P = 0.0083$, two-tailed permutation test with 9999 iterations using the absolute difference in the median frequencies as test statistic; Manly 2006).

Taken together, our analyses make clear that adaptive differentiation between basic and acidic stickleback ecotypes is built on the selection of genetic variation preexisting in the ancestor (i.e., standing genetic variation)—all core SNP alleles found in freshwater were also present in the sea. For the two core SNP regions scrutinized by targeted sequencing, the repeated use of standing variation during parallel evolution was confirmed directly by haplotype sharing among similar ecotypes from different lakes. Furthermore, adaptation to the ecologically relatively extreme acidic lakes has involved the accumulation of genetic variants relatively uncommon in marine fish, whereas the basic ecotypes have mostly retained the allele predominant in the sea. This B-A asymmetry in marine core SNP allele frequencies provides a strong indication that these regions are truly involved in adaptation.

An intriguing question is why acidic core SNP alleles still occur at relatively appreciable frequencies in the sea (Fig. 4C left)—given that they represent polymorphisms tightly linked to genetic variants beneficial in an ecologically very different habitat type, and in part likely even coincide with such variants. A first possibility is that the acidic alleles are recessive and hence deleterious in the sea only when homozygous, thus impeding their complete elimination by selection (e.g., Cresko et al. 2004). The observed frequencies of most acidic core SNP alleles in the sea, however, seem too high for this scenario. Another explanation is that these frequencies reflect an antagonism between purifying selection in the marine population and gene flow from the acidic ecotypes (i.e., migration-selection balance), maintained by continued hybridization between acidic and marine stickleback. At first sight, this scenario may appear plausible, as marine stickleback are reported to migrate into coastal lagoons and some freshwater lakes on North Uist during the breeding period. However, hybridization between acidic and marine stickleback seems extremely rare (A.D.C. MacColl, personal observation). Fortunately, our data allow a more direct evaluation of the above migration-selection balance scenario: if gene flow between acidic and marine stickleback was common, we should find acidic alleles at higher frequency in our marine sample taken close to the drainages of the acidic lakes (OBSM on the east side of North Uist; Fig. 1A) than in the sample taken near the drainages of the basic lakes (ARDH, west side). Interestingly, this prediction is not upheld; the frequency of the acidic alleles at the core SNPs did not differ between the two marine samples separated by hundreds of kilometers of shoreline (Fig. 5). The frequency in the sea of alleles important to adaptive B-A differentiation is therefore not substantially influenced in the short term by gene flow from freshwater ecotypes. Instead, these frequencies

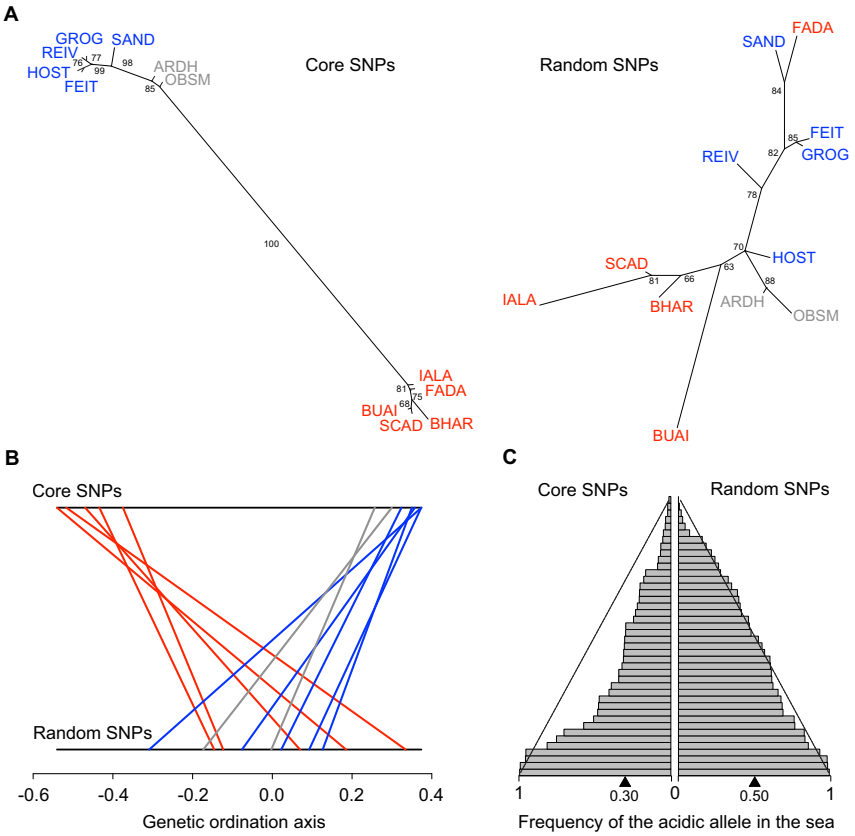


Figure 4. **(A) Unrooted phylogenies based on one synthetic individual per population, generated by drawing alleles at random at the core SNPs representing 42 regions of strong basic-acidic differentiation (left), and at 42 random SNPs (right). Population color codes follow Figure 1. Note the strong bootstrap support for the basal branches in the core SNP tree only. (B) Ordination (NMDS) of the populations, quantifying their genetic similarity across the 42 core (top) and random SNPs (bottom). Each line connects the position of a single population on the two coordinates. (C) Frequency distribution of the acidic allele in the marine stickleback across the core (left) and random SNPs (right). Within each category, the SNPs are shown in rows ordered by increasing frequency, and the black triangle on the bottom indicates their median frequency. The black dashed lines indicate allele frequencies expected under the uniform distribution.**

seem characteristic of marine stickleback around North Uist *in general*.

As a potential alternative explanation to migration-selection balance, we consider that alleles selected to high frequency in acidic lakes may not be deleterious within the marine habitat when segregating at modest frequency—a prediction from models of local adaptation involving polygenic traits (Latta 1998; Le Corre and Kremer 2012)—thus preventing their complete elimination in the sea. However, we cannot rule out the possibility that we systematically overestimate the marine frequencies of the actual variants favored in the acidic lakes, given that the physical linkage between these variants and the corresponding acidic core SNPs may not be perfect. Evaluating these different ideas will benefit from individual-level whole-genome sequence data from freshwater and marine stickleback on and around North Uist, and from direct information on the phenotypic role and fitness consequences of acidic freshwater alleles in the marine habitat.

THE RISE OF UNCOMMON ALLELES HAS CAUSED MORE NUMEROUS SELECTIVE SWEEPS IN THE ACIDIC ECOTYPES

We have demonstrated that the differentiation between basic and acidic stickleback ecotypes on North Uist has generally involved the retention of alleles common in the ancestor in the basic lakes,

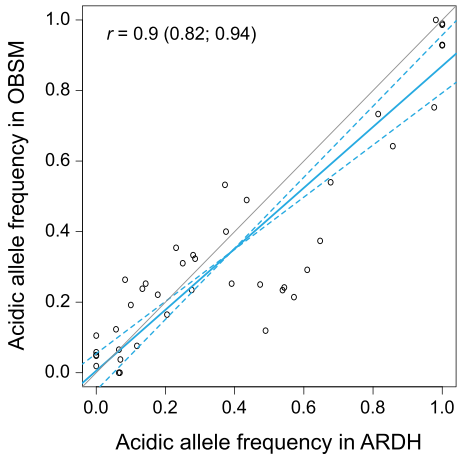


Figure 5. Association of the frequency of the acidic allele at the 42 core SNPs between the two marine stickleback samples (OBSM and ARDH). The slope of a major axis regression (blue line, with 95% confidence interval shown as dotted blue lines) is close to unity (gray line), indicating that the acidic alleles occur at similar frequencies in both samples. The regression statistic and associated 95% bootstrap CI are presented inside the graphic.

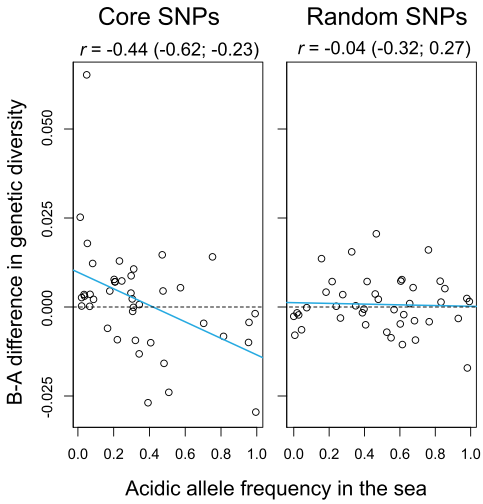


Figure 6. Selective sweeps in genomic regions important to B-A differentiation. The difference in genetic diversity (relative density of high-MAF SNPs) between basic and acidic populations across the 40 kb surrounding a focal SNP is plotted against the frequency of the acidic allele in the marine fish at the corresponding core (left) and random SNP (right). Each data point represents one of the 42 SNPs in each category. The statistics are Pearson's correlation coefficient along with its 95% bootstrap CI, and the blue lines show linear regressions (excluding the top left high-residual core SNP had a minimal influence on the relationship: $r = -0.45$, 95% CI −0.66 to −0.18).

and the selection of alleles uncommon in the ancestor in the acidic lakes. This implies that at loci important to B-A differentiation, the basic populations must mostly have experienced relatively weak allele frequency changes, or no changes at all, whereas the acidic populations must have experienced stronger allele frequency changes and hence stronger associated reductions in genetic diversity (selective sweeps; Maynard Smith and Haigh et al. 1974; Kaplan et al. 1989; Hermisson and Pennings 2005; Messer and Petrov 2013). Our inspection of genetic diversity, quantified by the relative density of high-MAF polymorphisms, across the 40 kb surrounding the core SNPs clearly supports this idea: the magnitude to which genetic diversity around a core SNP was reduced in acidic relative to basic stickleback was negatively correlated with the frequency of the corresponding acidic allele in the marine fish (Fig. 6 left). In other words, adaptation to the acidic lakes produced the strongest selective sweeps (positive B-A difference in genetic diversity) around those acidic variants segregating at the lowest frequency in the sea. Conversely, around the few core SNPs at which the acidic allele was the *predominant* one in the sea, strong allele frequency changes and associated selective sweeps tended to occur within the *basic* populations (negative B-A difference in genetic diversity). These observations offer a further validation of the ecological importance of the genomic regions tagged by our core SNPs. In addition, this analysis indicates that allele frequencies observed in present-day marine stickleback must be similar enough to those in the true marine ancestor of our freshwater populations to still allow detecting their association with patterns of genetic diversity shaped during adaptation.

At the random SNPs, we found no clear relationship between the frequency of the acidic alleles in the sea and habitat-related bias in genetic diversity (Fig. 6 right), as expected for polymorphisms neutral to B-A ecology. However, inspecting genetic diversity across an extended (1 Mb) segment around the random SNPs revealed interesting habitat-related patterns: the acidic populations tended to harbor lower diversity than the basic ones (two-tailed permutation test using population medians as data points and the B-A median difference as test statistic: $P = 0.0454$), and the highest diversity occurred in the marine fish (Fig. 7). The latter observation conforms to the common trend of marine stickleback to exhibit large effective population sizes—allowing the maintenance of elevated genetic diversity—relative to derived freshwater populations (Mäkinen et al. 2006; Hohenlohe et al. 2010; Catchen et al. 2013). The finding of low genetic diversity in the acidic ecotypes, however, seems surprising: the three largest lakes in our study are acidic

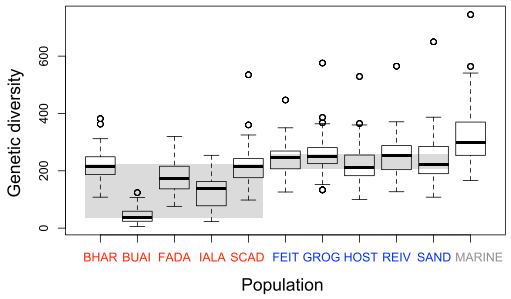


Figure 7. **Genetic diversity within populations, expressed as the number of SNPs passing a MAF threshold of 0.3 across 1 Mb windows centered at the 42 random SNPs, visualized by standard boxplots (i.e., thick lines represent the medians, rectangular boxes the interquartile ranges, IQR). The populations are ordered by habitat (acidic, basic, and marine). The gray background rectangles indicate the range of the medians across all five populations within the basic and within the acidic ecotypes. The marine population reflects the combination of the OBSM and ARDH samples; considering the larger of the two (OBSM, *N* = 20) alone, however, leads to very similar results (OBSM only: median diversity = 294, IQR = 251–356; OBSM and ARDH combined: median = 299, IQR = 254–370).**

(Fig. 1A), which would lead to the expectation of larger effective population sizes in the acidic than the basic lakes on average, and hence relatively reduced genetic diversity in the latter. However, our environmental and phenotypic data and our analyses of core SNP alleles consistently indicate that acidic lakes are ecologically more different from the ancestral habitat than the basic lakes. Ancestral colonizers must therefore have been exposed to particularly intense selection (i.e., been more strongly maladapted) within the acidic lakes, implying an initial period of particularly low population density. Moreover, acidic lakes display lower productivity than basic lakes (Waterston et al. 1979) and may therefore support relatively reduced population densities even in the long term. Both of these conditions would have promoted the stochastic loss of genetic variation in the acidic ecotypes. Clearly, our derived freshwater populations, and especially the acidic ones, lost genetic diversity not only within localized regions around targets of selection, but also genome-wide due to habitat-related differences in effective population size.

Conclusions

Our study shows that the emergence of similar ecotypes within multiple derived habitats can result in parallel evolution at the genomic level. We further demonstrate how insights into the differentiation of derived populations can be strengthened by including genetic data from their recent common ancestor: in our stickleback system, basic-acidic differentiation occurred via the genome-wide sorting of standing variation in the ancestor, and asymmetry in this sorting is predictable from the difference of each derived habitat from the ancestral one. The detection of numerous genomic regions repeatedly involved in basic-acidic differentiation now provides a resource for identifying the associated phenotypes and exploring their ecological function. Such work may reveal whether genomic regions showing the strongest parallelism include developmental components of bony armor traits, or if they represent more elusive aspects of adaptive differentiation between basic and acidic waters.

AUTHOR CONTRIBUTIONS

D.B., M.R., and Q.H. conceived the study; A.D.C.M. performed field sampling and measurements; D.M., M.R., D.B., and Q.H. performed wet lab work; Q.H. and D.B. analyzed and interpreted data; Q.H. and D.B. wrote the manuscript, with input from all co-authors.

ACKNOWLEDGMENTS

We thank North Uist Estates, the Scottish Government (SEERAD) and the North Uist Angling Club for access to the lakes on North Uist; the Swiss National Science Foundation (SNF; grants 31003A_146208 and 31003A_165826) and the Freiwillige Akademische Gesellschaft Basel (FAG) for financial support to DB; the Natural Environment Research Council (NERC; grant NE/J02239X/1) for financial support to ADCM; Daniele D'Agostino for assisting with sampling; Walter Salzburger for sharing wet lab infrastructure; Brigitte Aeschbach and Nicolas Boileau for facilitating lab work; Christian Beisel, Ina Nissen, and Elodie Burcklen for sequencing at the Quantitative Genomics Facility, D-BSSE, ETH Zürich; Thierry Sengstag for organizing access to the BC2 cluster; the developers of Novocraft for making their aligner freely available; Laurent Guerard for help with scripting; Anja Frey, Telma G. Laurentino, Salome Hosch, and Rike Teuber for contributing to analysis and/or laboratory work; Mirjam Bissegger, Gleb Ebert, and Fabrizia Ronco for phenotyping; Graham Coop and six anonymous reviewers for valuable suggestions to improve this article.

DATA ARCHIVING

All raw Illumina sequences, demultiplexed by population (freshwater samples) or by individual (marine samples) are available from the NCBI Sequence Read Archive under BioProject number PRJNA485717. Key datasets and R codes are available from Dryad (https://doi.org/10.5061/dryad.4ck2q0m)

CONFLICT OF INTEREST

The authors declare no conflicts of interest.

LITERATURE CITED

Arendt, J., and D. Reznick. 2008. Convergence and parallelism reconsidered: what have we learned about the genetics of adaptation? Trends Ecol. Evol. 23:26–32.

Baird, N. A., P. D. Etter, T. S. Atwood, M. C. Currey, A. L. Shiver, Z. A. Lewis, et al. 2008. Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS One 3:e3376.

Ballantyne, C. 2010. Extent and deglacial chronology of the last British-Irish Ice Sheet: implications of exposure dating using cosmogenic isotopes. J. Quaternary. Sci. 25:515–534.

Bell, M. F., and S. Foster. 1994. The evolutionary biology of the threespine stickleback. Oxford Univ. Press, Oxford, U.K.

Berner, D., and M. Roesti. 2017. Genomics of adaptive divergence with chromosome-scale heterogeneity in crossover rate. Mol. Ecol. 26:6351–6369.

Berner, D., and W. Salzburger. 2015. The genomics of organismal diversification illuminated by adaptive radiations. Trends Genet. 31:491–499.

Bierne, N., P. A. Gagnaire, and P. David. 2013. The geography of introgression in a patchy environment and the thorn in the side of ecological speciation. Curr. Zool. 59:72–86.

Braverman, J. M., R. R. Hudson, N. L. Kaplan, C. H. Langley, and W. Stephan. 1995. The hitchhiking effect on the site frequency-spectrum of DNA polymorphisms. Genetics 140:783–796.

Campbell, R. N., and R. B. Williamson. 1979. The fishes of inland waters in the Outer Hebrides. Proc. Roy. Soc. Edinb. B. 77:377–393.

Catchen, J., S. Bassham, T. Wilson, M. Currey, C. O'Brien, Q. Yeates, et al. 2013. The population structure and recent colonization history of Oregon threespine stickleback determined using restriction-site associated DNA-sequencing. Mol. Ecol. 22(11):2864–2883.

Cresko, W. A., A. Amores, C. Wilson, J. Murphy, M. Currey, P. Phillips, et al. 2004. Parallel genetic basis for repeated evolution of armor loss in Alaskan threespine stickleback populations. Proc. Natl. Acad. Sci. USA 101:6050–6055.

El Nagar, A., and A. D. C. MacColl. 2016. Parasites contribute to ecologically dependent postmating isolation in the adaptive radiation of three-spined stickleback. Proc. Roy. Soc. B-Biol. Sci. 283:20160691.

Elgvin, T. O., C. N. Trier, O. K. Torresen, I. J. Hagen, S. Lien, A. J. Nederbragt, et al. 2017. The genomic mosaicism of hybrid speciation. Sci. Adv. 3:e1602996.

Endler, J. A. 1986. Natural selection in the wild. Princeton Univ. Press, Princeton, NJ.

Gautier, M. 2015. Genome-wide scans for adaptive differentiation and association analysis with population-specific covariables. Genetics 201:1555–1579

Giles, N. 1983. The possible role of environmental calcium levels during the evolution of phenotypic diversity in Outer Hebridean populations of the 3-spined stickleback, *Gasterosteus aculeatus*. J. Zool. 199:535–544.

Glazer, A. M., E. E. Killingbeck, T. Mitros, D. S. Rokhsar, and C. T. Miller. 2015. Genome assembly improvement and mapping convergently evolved skeletal traits in sticklebacks with genotyping-by-Sequencing. G3-Genes. Genom. Genet. 5:1463–1472.

Haenel, Q., T. G. Laurentino, M. Roesti, and D. Berner. 2018. Meta-analysis of chromosome-scale crossover rate variation in eukaryotes and its significance to evolutionary genomics. Mol. Ecol. 27:2477–2497.

Hermisson, J. and P. S. Pennings. 2005. Soft sweeps: molecular population genetics of adaptation from standing genetic variation. Genetics 169:2335–2352.

Hohenlohe, P. A., S. Bassham, P. D. Etter, N. Stiffler, E. A. Johnson and W. A. Cresko. 2010. Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. Plos. Genet. 6:e1000862.

Jones, F. C., Y. F. Chan, J. Schmutz, J. Grimwood, S. D. Brady, A. M. Southwick, et al. 2012a. A genome-wide SNP genotyping array reveals patterns of global and repeated species-pair divergence in sticklebacks. Curr. Biol. 22:83–90.

Jones, F. C., M. G. Grabherr, Y. F. Chan, P. Russell, E. Mauceli, J. Johnson, et al. 2012b. The genomic basis of adaptive evolution in threespine sticklebacks. Nature 484:55–61.

Kaplan, N. L., R. R. Hudson and C. H. Langley. 1989. The "hitchhiking effect" revisited. Genetics 123:887–899.

Kawasaki, K. 2009. The SCPP gene repertoire in bony vertebrates and graded differences in mineralized tissues. Dev. Genes. Evol. 219:147–157.

Kawasaki, K., T. Suzuki, and K. M. Weiss. 2004. Genetic basis for the evolution of vertebrate mineralized tissue. Proc. Natl. Acad. Sci. USA 101:11356–11361.

Klepaker, T., K. Østbye, R. Spence, M. Warren, M. Przybylski, and C. Smith. 2016. Selective agents in the adaptive radiation of Hebridean sticklebacks. Evol. Ecol. Res. 17:243–262.

Lamichhaney, S., F. Han, J. Berglund, C. Wang, M. S. Almen, M. T. Webster, et al. 2016. A beak size locus in Darwin's finches facilitated character displacement during a drought. Science 352:470–474.

Latta, R. G. 1998. Differentiation of allelic frequencies at quantitative trait loci affecting locally adaptive traits. Am. Nat. 151:283–292.

Le Corre, V., and A. Kremer. 2012. The genetic differentiation at quantitative trait loci under local adaptation. Mol. Ecol. 21:1548–1566.

Lescak, E. A., S. L. Bassham, J. Catchen, O. Gelmond, M. L. Sherbick, F. A. von Hippel, et al. 2015. Evolution of stickleback in 50 years on earthquake-uplifted islands. Proc. Natl. Acad. Sci. USA 112:E7204–E7212.

Lowry, D. B., S. Hoban, J. L. Kelley, K. E. Lotterhos, L. K. Reed, M. F. Antolin, et al. 2017. Breaking RAD: an evaluation of the utility of restriction site-associated DNA sequencing for genome scans of adaptation. Mol. Ecol. Resour. 17:142–152.

MacColl, A. D. C., A. El Nagar, and J. de Roij. 2013. The evolutionary ecology of dwarfism in three-spined sticklebacks. J. Anim. Ecol. 82:642–652.

Magalhaes, I. S., D. D. Agostino, P. A. Hohenlohe, and A. D. C. Maccoll. 2016. The ecology of an adaptive radiation of three-spined stickleback from North Uist, Scotland. Mol. Ecol. 25:4319–4336.

Mäkinen, H. S., J. M. Cano, and J. Merilä. 2006. Genetic relationships among marine and freshwater populations of the European three-spined stickleback (*Gasterosteus aculeatus*) revealed by microsatellites. Mol. Ecol. 15:1519–1534.

Manly, B. F. J. 2006. Randomization, bootstrap and Monte-Carlo methods in biology. 3rd Edition. Chapman and Hall, Boca Raton, FL.

Marques, D. A., J. S. Taylor, F. C. Jones, F. Di Palma, D. M. Kingsley, and T. E. Reimchen. 2017. Convergent evolution of SWS2 opsin facilitates adaptive radiation of threespine stickleback into different light environments. PLoS Biol. 15:e2001627.

Martin, S. H., K. K. Dasmahapatra, N. J. Nadeau, C. Salazar, J. R. Walters, F. Simpson, et al. 2013. Genome-wide evidence for speciation with gene flow in Heliconius butterflies. Genome. Res. 23:1817–1828.

Matschiner, M. 2016. Fitchi: haplotype genealogy graphs based on the Fitch algorithm. Bioinformatics 32:1250–1252.

Maynard Smith, J., and J. Haigh. 1974. The hitch-hiking effect of a favourable gene. Genet. Res. 23:23–35.

Messer, P. W., and D. A. Petrov. 2013. Population genomics of rapid adaptation by soft selective sweeps. Trends Ecol. Evol. 28:659–669.

Morgan, M., H. Pagès, V. Obenchain, and N. Hayden. 2017. Rsamtools: Binary alignment (BAM), FASTA, variant call (BCF), and tabix file import. R package version 1.30.0.

R Core Team. 2017. R: A language and environment for statistical computing, Austria. https://www.R-project.org/

Reid, N. M., D. A. Proestou, B. W. Clark, W. C. Warren, J. K. Colbourne, J. R. Shaw, et al. 2016. The genomic landscape of rapid repeated evolutionary adaptation to toxic pollution in wild fish. Science 354:1305–1308.

Roesti, M., S. Gavrilets, A. P. Hendry, W. Salzburger, and D. Berner. 2014. The genomic signature of parallel adaptation from shared genetic variation. Mol. Ecol. 23:3944–3956.

Roesti, M., B. Kueng, D. Moser, and D. Berner. 2015. The genomics of ecological vicariance in threespine stickleback fish. Nat. Commun. 6:10229.

Roesti, M., D. Moser, and D. Berner. 2013. Recombination in the threespine stickleback genome–patterns and consequences. Mol. Ecol. 22:3014–3027.

Roesti, M., W. Salzburger, and D. Berner. 2012. Uninformative polymorphisms bias genome scans for signatures of selection. BMC Evol. Biol. 12:94.

Schliep, K. P. 2011. phangorn: phylogenetic analysis in R. Bioinformatics 27:592–593.

Schluter, D. 2000. The ecology of adaptive radiation. Oxford Univ. Press, Oxford, U.K.

Spence, R., R. J. Wootton, I. Barber, M. Przybylski, and C. Smith. 2013. Ecological causes of morphological evolution in the three-spined stickleback. Ecol. Evol. 3:1717–1726.

Stamatakis, A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313.

Stephens, M., N. J. Smith, and P. Donnelly. 2001. A new statistical method for haplotype reconstruction from population data. Am. J. Hum. Genet. 68:978–989.

Terekhanova, N. V., M. D. Logacheva, A. A. Penin, T. V. Neretina, A. E. Barmintseva, G. A. Bazykin, et al. 2014. Fast evolution from precast bricks: genomics of young freshwater populations of threespine stickleback *Gasterosteus aculeatus*. PLoS Genet. 10:e1004696.

Waterston, A. H., A. V. Holden, R. N. Campbell, and P. Maitland. 1979. The inland waters of the Outer Hebrides. Proc. Roy. Soc. Edinb. B. 77:329–351.

Wright, S. 1931. Evolution in Medelian populations. Genetics 16:97–159.

Yeaman, S., K. A. Hodgins, K. E. Lotterhos, H. Suren, S. Nadeau, J. C. Degner, et al. 2016. Convergent local adaptation to climate in distantly related conifers. Science 353:1431–1433.

Associate Editor: Z. Gompert

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Method.
Discussion.
Table S1. Description of study habitats and populations.
Table S2. All pairwise populations comparisons (AFD and F_{ST}).
Table S3. Pairwise comparisons per habitat (AFD and F_{ST}).
Table S4. List of genes around the top core SNPs.
Table S5. Characterization of the 42 core SNPs.
Figure S1. Schematic description of the SNP generation protocol.
Figure S2. Unrooted nuclear phylogeny (neutral SNPs).
Figure S3. Marine-freshwater differentiation by chromosome.
Figure S4. Standardized basic-acidic differentiation.
Figure S5. BayPass analysis.
Figure S6. Robustness check for random SNPs.
Figure S7. Comparision SNP density vs. nucleotide diversity (π).
Figure S8. Robustness check for selective sweeps.
Figure S9. Unrooted nuclear phylogeny (NJ tree, neutral SNPs).
Figure S10. Unrooted nuclear phylogeny (all loner SNPs).
Figure S11. Unrooted nuclear phylogeny (SNP spacing at least 1Mb).
Figure S12. Description of the top core SNPs.
Figure S13. Genome-wide basic-acidic differentiation.
Figure S14. Potential chromosomal inversions.

Supporting Information

Predictable genome-wide sorting of standing genetic variation during parallel adaptation to basic versus acidic environments in stickleback fish

Quiterie Haenel[1], Marius Roesti[1,2,3], Dario Moser[1,4], Andrew D. C. MacColl[5] and Daniel Berner[1]

[1] Department of Environmental Sciences, Zoology, University of Basel, 4051 Basel, Switzerland

[2] Biodiversity Research Centre and Zoology Department, University of British Columbia, Vancouver, British Columbia V6T 1Z4, Canada

[3] *Current address*: Institute of Ecology and Evolution, University of Bern, 3012 Bern, Switzerland

[4] *Current address*: Jagd- und Fischereiverwaltung Thurgau, 8510 Frauenfeld, Switzerland

[5] School of Life Sciences, University of Nottingham, Nottingham NG7 2RD, United Kingdom

Corresponding authors:

- Quiterie Haenel: quiterie.haenel@unibas.ch
- Daniel Berner: daniel.berner@unibas.ch

1

Contents

Methods

Targeted individual-level Sanger sequencing was performed at two top AFD extremes by amplifying a 700 bp fragment from a subsample of 4-8 individuals from each of the 12 populations. Primer pairs and PCR conditions were as follows:

Chromosome IX, SNP position (bp): 13,360,688

Primer	Sequence
Forward	5'CAGTCAGAGGACCGGACGT3'
Reverse	5'ATCTCTGCTGATGGTTGGCA3'

For a 12.5uL reaction volume, we used 1.25uL Taq polymerase buffer (x10), 1uL dNTP mix (final concentration of each dNTP 200uL), 0.25uL of each primer at 10uL, 1uL of DNA template, 0.50uL of Red Taq DNA polymerase and 8.25uL of sterile deionized water. Cycling conditions were 2 min at 94°C (1 cycle); 30 sec at 94°C, 30 sec at 60°C, 1 min at 72°C (30 cycles); 7 min at 72°C (1 cycle). PCR success was confirmed on a 1.5% agarose gel.

Chromosome IV, SNP position (bp): 26,641,811

Primer	Sequence
Forward	5'AGCCACAATGCCAAAGGACA3'
Reverse	5'CAAATCCAAACACTCGGGTGG3'

For a 12.5uL reaction volume, we used 1.25uL Taq polymerase buffer (x10), 1uL dNTP mix (final concentration of each dNTP 200uL), 0.25uL of each primer at 10uL,

1uL of DNA template, 0.50uL of Red Taq DNA polymerase, 0.75uL of MgCl2 and 7.5uL of sterile deionized water. Cycling conditions were 2 min at 94°C (1 cycle); 30 sec at 94°C, 30 sec at 54°C, 1 min at 72°C (30 cycles); 7 min at 72°C (1 cycle). PCR success was confirmed on a 1.5% agarose gel.

4

Discussion

This discussion presents additional detail and evidence supporting conclusions drawn in the main paper.

1) Is it valid to combine the two marine samples to a single marine population in the present study? Would exploring the evolutionary independence of the derived freshwater populations not require including samples from additional marine populations?

Very generally, marine stickleback occurring in a broader geographic region are considered a large, genetically well mixed population; they display very limited genetic structure compared to derived freshwater populations from the same region, and elevated genetic diversity relative to freshwater populations (Hohenlohe et al. 2010; Jones et al. 2012a; Catchen et al. 2013; Roesti et al. 2014). As expected, these classical genetic patterns are also observed in the present study: although our two marine samples were taken from sites separated by more than a hundred kilometers of shoreline (Fig. 1A), their comparison yields a median genome- wide differentiation of only 0.07 (AFD) and 0.01 (F_{ST}) (Table S2 and S3; note that this magnitude of genetic differentiation is almost certainly overestimated because the marine samples were substantially smaller [N = 10 and 20] than all the freshwater samples [Table S1], and the associated imprecision in allele frequency estimation should bias both median AFD and F_{ST} upward). By contrast, median genome-wide differentiation averaged across all comparisons within each freshwater habitat type (both N = 10) is much greater (basic-basic comparisons: AFD = 0.17, F_{ST} = 0.04; acidic-acid comparisons: AFD = 0.25, F_{ST} = 0.09), despite the populations within

5

each freshwater habitat type being separated by much smaller geographic distances. Moreover, Fig. 7 reveals greater genetic diversity in the marine than the freshwater fish, consistent with the generally large effective population size of marine stickleback.

While these observed patterns of genetic differentiation and diversity are fully consistent with work on marine and freshwater stickleback worldwide and indicate that marine stickleback around North Uist can be considered a large, well-mixed population, even more compelling evidence emerges from our phylogenies: in all trees (Fig. 2A, Fig S2, S9, S10, S11), the branches connecting each marine sample (OBSM, ARDH) to their first common node are very similar in length. This means that no marine sample can be considered closer to any of the freshwater populations than the other marine sample (a similar pattern emerges when exploring population similarity by ordination, Fig. 2B). This in turn implies that even a single marine sample would provide a sufficient proxy of the marine ancestor of all the derived freshwater populations. Clearly, combining our two marine samples to a single biological population is a valid approach; additional marine samples are not needed for our analyses.

2) The generation of synthetic individuals based on pooled sequencing genotype data generates artificial linkage equilibrium among alleles – could this bias phylogenetic inference in the present study?

As a robustness check of using synthetic individuals for phylogenetic inference, we repeated the phylogenetic analysis using individuals generated by concatenating nucleotides from SNPs spaced by a minimum of 1 Mb only. Since linkage

6

disequilibrium has been observed in threespine stickleback to decay over a physical distance of a few kilobases (e.g., Roesti et al. 2015), this spacing should ensure that concatenated alleles can also occur on the same DNA molecule in nature. Despite limited marker resolution (227 SNPs only), this alternative phylogenetic analysis (Fig. S11) recovered the main features of the high-resolution trees.

We recognize, however, that the concatenation of nucleotides from a pool may be problematic when linkage disequilibrium within populations is strong over large physical scales. The main scenario able to generate such linkage disequilibrium is recent dispersal among populations. In this case, long immigrant haplotype tracts differing from the standard genetic composition of a given population would be disintegrated during DNA pooling so that synthetic individuals derived from the pooled sequence data would appear more similar in the phylogenetic tree than would real individual genotypes. A scenario of recent dispersal among populations, however, can be ruled out for our study: first, all our populations show strong pairwise genetic differentiation from each other (Table S2 and S3; see also previous paragraph). Second, given the present-day hydrology of the study system, only marine-freshwater dispersal would be plausible. However, marine fish are phenotypically distinct from the basic and acidic ecotypes, so that marine-freshwater migrants (and likely even recent hybrids and backcrosses) could be identified phenotypically. Our phenotypic analysis, however, yielded no indication of migration or hybridization. Third, a recent study using individual-level sequence data (Magalhaes et al. 2016), covering seven out of our ten freshwater populations, found no indication of population admixture (Fig. 3 in that paper). We therefore see no reason to assume long-range linkage disequilibrium within our populations, and are

confident that our phylogenetic analysis using nucleotide concatenation produces reliable insights into the genetic similarity among our study populations. In support of this view, our general observation of population monophyly is consistent with monophyly observed in a tree based on individual-level genotype data from a subset of our study populations (Fig. S1 in Magalhaes et al. 2016).

3) Can the study rule out the possibility that each of the two freshwater stickleback ecotypes (basic and acidic) evolved only once on North Uist, expanded geographically, came into secondary contact, and started hybridizing? The resulting genetic exchange may have caused some basic and acidic populations to cluster together on the terminal branches of the genealogical tree, thus falsely suggesting the repeated independent differentiation of basic and acidic populations (Bierne et al. 2013).

This possibility appears extremely unparsimonious when interpreting our phylogenetic trees (Fig. 2A, S2 and S9, S10, S11) in the light of the geographic arrangement of the lakes and habitat types. Specifically, because of the geologically determined (Waterson et al. 1979; Giles 1983) spatial segregation between the two habitat types (basic in the west and acidic in the east; Fig. 1A), it does not appear physically and ecologically plausible that an ancient acidic ecotype dispersed to the basic region and vice versa. The basic and acidic catchments are widely separated in space, and the only aquatic route between them is through the sea. In addition, the specific habitat appropriate to each ecotype would have been missing in the newly invaded region, making successful dispersal highly unlikely. Secondary contact and introgressive genetic exchange between the ecotypes across the entire island is

therefore not realistic for geological and ecological reasons. The relative genetic similarity between, for example, the HOST and IALA populations (observed consistently in all our genealogies, see Fig. 2A, S2 and S9, S10, S11) can only be explained plausibly by *independent* colonization from a large, genetically well-mixed marine population, followed by the stochastic sorting of ancestral neutral variation that has resulted in these populations being relatively similar genetically by chance. Moreover, if dispersal and gene flow had been extensive at the scale of the entire island, it would be hard to explain why populations of the *same* ecotype and residing in close geographic neighborhood (e.g., the IALA and BUAI populations) consistently emerge as genetically distant in all our phylogenies, and also in the ordination (Fig. 2B). The repeated stochastic sorting of neutral genetic variation from a shared marine ancestor during independent evolution is the only explanation parsimoniously reconciling our tree topology with the geography of the study populations. Our tree- and ordination-based inference of evolutionary independence is also supported by the general absence of substantial allele frequency correlation between populations, as estimated by BayPass (Fig. S5B), and fully consistent with our study lakes currently draining independently into the sea (not confirmed for FEIT; Fig. 1A).

Further evidence of the repeated, independent evolution of similar ecotypes in multiple lakes derives from the non-perfect phenotypic parallelism among the acidic populations (Fig. 1B, Table S1): the IALA population, for instance, exhibits a fully developed pelvic structure like the basic ecotype, whereas the FADA population in very close neighborhood (Fig. 1A) has completely lost its pelvic structure. Such genetically based phenotypic differences are difficult to explain when assuming the formation and spread of a single ancestral acidic ecotype across the acidic side of

North Uist. Conversely, the relative similarity of the basic populations (Fig. 1B, Table S1) does not imply that they derive from a single ancestral basic ecotype that emerged once on North Uist; the basic populations correspond phenotypically to the standard freshwater stickleback ecotype known to have evolved independently through parallel differentiation from marine ancestors countless times all across the species' range (Bell & Foster 1994).

Tables

Table S1: Characterization of the lakes and lagoons from which the stickleback samples were collected, and number of individuals sampled from each site (numbers in parentheses indicate sample sizes underlying Sanger sequencing of each of the two loci in Fig. 3). Data on pH, water surface, and calcium concentration (Ca2+) and lake surface are from Magalhaes *et al.* (2016); this publication also provides geographic coordinates of all lakes. Armor trait data are averaged over 20 individuals chosen at random within each sample (with the exception of the marine samples that were considered in full). Lateral plate number refers to a single body side.

Habitat type	Site code	Site name	N	pH	[Ca^{2+}] (10^{-5} mg/L)	Water surface (ha)	Lateral plate number	Dorsal spines number	Presence of a pelvic complex (%)
Acidic	BHAR	a' Bharpa	30 (4/3)	6	3.42	53.9	0	1.85	0
Acidic	BUAI	na Buaile	30 (4/4)	6.7	6.01	1.7	2.95	3	30
Acidic	FADA	Fada	23 (4/4)	6.7	4.06	160.0	0	0.5	0
Acidic	IALA	Ialaidh	26 (4/4)	6.4	5.95	0.4	3.05	2.45	100
Acidic	SCAD	Scadavary	30 (4/2)	6.1	3.27	551.6	0.2	1.65	5
Basic	FEIT	nam Feithean	30 (3/4)	8.3	77.6	15.7	4.25	2.95	100
Basic	GROG	Grogary	30	8.2	63.8	14.8	2.95	2.95	100

			(4/3)						
Basic	HOST	Hosta	30 (4/4)	8.3	72.3	25.8	3.4	3	100
Basic	REIV	na Reival	29 (4/4)	9	44.9	6.1	3.4	3	100
Basic	SAND	Sandary	30 (3/4)	8.3	75.2	15.5	3	3	100
Marine	ARDH	Ard Heisker	10 (7/8)	8.6	498.5	-	25	3	100
Marine	OBSM	Ob' nan Stearnain	20 (7/8)	9.1	487.1	-	25	3	100

12

Table S2: Genome-wide mean (lower-left semimatrix) and median (upper-right semimatrix) genetic differentiation, expressed as absolute allele frequency difference (AFD), and as F_{ST} (Nei's 1973 estimator G_{ST}) in parentheses, for all pairwise populations comparisons.

	BHAR	BUAI	FADA	IALA	SCAD	FEIT	GROG	HOST	REIV	SAND	ARDH	OBSM
BHAR		0.344 (0.171)	0.219 (0.068)	0.246 (0.086)	0.160 (0.038)	0.207 (0.055)	0.218 (0.061)	0.214 (0.061)	0.210 (0.061)	0.205 (0.057)	0.196 (0.066)	0.192 (0.063)
BUAI	0.384 (0.252)		0.339 (0.169)	0.287 (0.132)	0.349 (0.175)	0.363 (0.184)	0.387 (0.202)	0.376 (0.195)	0.368 (0.192)	0.358 (0.182)	0.345 (0.204)	0.354 (0.211)
FADA	0.267 (0.133)	0.395 (0.275)		0.246 (0.088)	0.215 (0.068)	0.238 (0.076)	0.241 (0.078)	0.238 (0.077)	0.242 (0.082)	0.224 (0.071)	0.226 (0.089)	0.227 (0.086)
IALA	0.302 (0.164)	0.385 (0.277)	0.314 (0.184)		0.249 (0.088)	0.267 (0.096)	0.272 (0.099)	0.262 (0.094)	0.261 (0.097)	0.245 (0.086)	0.247 (0.105)	0.250 (0.107)
SCAD	0.206 (0.085)	0.387 (0.257)	0.264 (0.131)	0.304 (0.167)		0.211 (0.058)	0.219 (0.063)	0.217 (0.063)	0.219 (0.066)	0.200 (0.056)	0.200 (0.068)	0.195 (0.065)
FEIT	0.249 (0.109)	0.386 (0.240)	0.281 (0.138)	0.306 (0.159)	0.252 (0.112)		0.074 (0.008)	0.139 (0.027)	0.184 (0.045)	0.201 (0.052)	0.194 (0.059)	0.191 (0.056)
GROG	0.262 (0.119)	0.412 (0.266)	0.287 (0.146)	0.314 (0.169)	0.263 (0.122)	0.090 (0.017)		0.135 (0.026)	0.192 (0.050)	0.207 (0.056)	0.207 (0.068)	0.205 (0.066)
HOST	0.261 (0.122)	0.406 (0.268)	0.289 (0.149)	0.311 (0.170)	0.265 (0.127)	0.176 (0.061)	0.171 (0.058)		0.200 (0.054)	0.210 (0.059)	0.197 (0.064)	0.194 (0.061)
REIV	0.256 (0.120)	0.400 (0.267)	0.291 (0.153)	0.312 (0.173)	0.267 (0.131)	0.223 (0.092)	0.234 (0.101)	0.244 (0.110)		0.212 (0.062)	0.176 (0.054)	0.177 (0.054)
SAND	0.255 (0.119)	0.394 (0.261)	0.277 (0.143)	0.296 (0.159)	0.251 (0.119)	0.242 (0.106)	0.252 (0.114)	0.257 (0.120)	0.259 (0.123)		0.196 (0.067)	0.194 (0.064)
ARDH	0.247 (0.150)	0.382 (0.282)	0.275 (0.171)	0.297 (0.189)	0.250 (0.152)	0.236 (0.133)	0.250 (0.148)	0.244 (0.143)	0.223 (0.132)	0.244 (0.150)		0.073 (0.010)
OBSM	0.243 (0.143)	0.390 (0.286)	0.274 (0.166)	0.300 (0.187)	0.247 (0.147)	0.233 (0.128)	0.247 (0.143)	0.241 (0.137)	0.224 (0.129)	0.242 (0.143)	0.096 (0.023)	

Table S3: Genome-wide mean and median AFD and F_{ST} (Nei's 1973 estimator G_{ST}) averaged across all population comparisons available for each habitat comparison category (A = acidic, B = basic, M = marine; number of population comparisons in parentheses).

Habitats	Mean		Median	
	AFD	**F_{ST}**	**AFD**	**F_{ST}**
A vs. A (10)	0.319	0.191	0.252	0.094
B vs. B (10)	0.216	0.091	0.165	0.038
B vs. A (25)	0.300	0.161	0.250	0.087
M vs. A (10)	0.289	0.185	0.234	0.097
M vs. B (10)	0.238	1.137	0.192	0.060
M vs. FW (20)	0.262	0.160	0.210	0.075
M vs. M (1)	0.096	0.023	0.073	0.010

Table S4: List of all the genes present in a 100 kb window around 19 core SNPs passing the stringent AFD threshold of 0.75 (i.e., the genome regions A to S characterized in Fig. S8).

ID	Chr	N°	Gene ID	Name	Start position	End position	Description
A	chrIX	1	ENSGACG00000017898	*odam*	13376375	13378298	odontogenic, ameloblast asssociated
		2	ENSGACG00000017900	*CNGA1*	13331902	13335265	cyclic nucleotide gated channel alpha 1
		3	ENSGACG00000017892	*sparcl1*	13386748	13392238	SPARC-like 1
		4	ENSGACG00000017889		13394935	13397012	osteopontin domain
		5	ENSGACG00000017903	*TACR3*	13314220	13326105	tachykinin receptor 3
		6	ENSGACG00000017887	*aptx*	13398416	13400992	aprataxin
		7	ENSGACG00000017879	*dnaja1*	13401116	13407421	DnaJ (Hsp40) homolog, subfamily A, member 1
		8	ENSGACG00000017872	*smu1b*	13410460	13415580	smu-1 suppressor of mec-8 and unc-52 homolog b (C. elegans)

B	chrI	1	ENSGACG00000004934	*supt5h*	866290	879634	SPT5 homolog, DSIF elongation factor subunit
		2	ENSGACG00000004963	*cox7a1*	885557	886553	cytochrome c oxidase subunit VIIa polypeptide 1 (muscle)
		3	ENSGACG00000004964	*nf1a*	889314	922850	neurofibromin 1a
		4	ENSGACG00000004929	*triap1*	864993	865217	TP53 regulated inhibitor of apoptosis 1
		5	ENSGACG00000004927		846012	847417	
		6	ENSGACG00000004922		835507	842787	
		7	ENSGACG00000004992	*smco4*	925191	925370	single-pass membrane protein with coiled-coil domains 4
C	chrIV	1	ENSGACG00000018959	*wnt7b*	26620620	26626229	wingless-type MMTV integration site family, member 7Ba
		2	ENSGACG00000018960	*atxn10*	26609883	26617247	ataxin 10
		3	ENSGACG00000018958	*pparaa*	26666495	26676255	peroxisome proliferator-activated receptor alpha a

16

		4	ENSGACG00000018964	*FBLN1*	26571948	26609189	fibulin 1
		5	ENSGACG00000018957	*si:ch211-239e6.4*	26680344	26681228	cysteine rich DPF motif domain containing 1
D	chrVII	1	ENSGACG00000020121	*lim2.1*	13813116	13815236	lens intrinsic membrane protein 2.1
		2	ENSGACG00000020120	*bsx*	13808290	13809957	brain-specific homeobox
		3	ENSGACG00000020119		13803433	13805401	
		4	ENSGACG00000020118		13799154	13803100	
		5	ENSGACG00000020117	*hspa8*	13782601	13786688	heat shock protein 8
		6	ENSGACG00000020116		13776594	13781629	
E	chrXX	1	ENSGACG00000007563		10619866	10623759	
		2	ENSGACG00000007569		10613890	10618424	
		3	ENSGACG00000007594	*zgc:171592*	10611768	10613631	chymotrypsin-like
		4	ENSGACG00000007597	*si:dkey-*	10607579	10608562	

17

				117a8.4			
		5	ENSGACG00000007546	*si:ch73-380l3.1*	10634711	10725421	
		6	ENSGACG00000007600	*lin37*	10602974	10605770	lin-37 DREAM MuvB core complex component
		7	ENSGACG00000007557		10637722	10708611	
		8	ENSGACG00000007618		10598096	10599752	
		9	ENSGACG00000007622	*hspb6*	10596751	10597728	heat shock protein, alpha-crystallin-related, b6
		10	ENSGACG00000007626	*psenen*	10593675	10594992	presenilin enhancer gamma secretase subunit
		11	ENSGACG00000007639		10573314	10585770	
		12	ENSGACG00000007659	*igflr1*	10565975	10570695	IGF-like family receptor 1
F	chrV	1	ENSGACG00000005578	*pemt*	3912342	3951075	phosphatidylethanolamine N-methyltransferase
		2	ENSGACG00000005572	*rasd1*	3956364	3957372	RAS, dexamethasone-induced 1
		3	ENSGACG00000005546	*NT5C*	3982133	3987560	5', 3'-nucleotidase, cytosolic

18

		4	ENSGACG00000005506	*cops3*	3987270	3995324	COP9 signalosome subunit 3
		5	ENSGACG00000005496	*usp22*	3996761	4006817	ubiquitin specific peptidase 22
G	chrIV	1	ENSGACG00000018803	*rassf8b*	28641481	28646130	Ras association (RalGDS/AF-6) domain family (N-terminal) member 8b
H	chrIV	1	ENSGACG00000018231	*abcb7*	12013745	12034920	ATP-binding cassette, sub-family B (MDR/TAP), member 7
		2	ENSGACG00000018229	*uprt*	12009666	12013671	uracil phosphoribosyltransferase (FUR1) homolog (S. cerevisiae)
		3	ENSGACG00000018224	*zdhhc15b*	12003229	12006996	zinc finger, DHHC-type containing 15b
I	chrXIV	1	ENSGACG00000017111	*pptc7b*	7257631	7262371	PTC7 protein phosphatase homolog b
		2	ENSGACG00000017108	*prnpb*	7255200	7256612	prion protein b
		3	ENSGACG00000017119	*aplnrb*	7265980	7266984	apelin receptor b
		4	ENSGACG00000017107		7252229	7252705	
		5	ENSGACG00000017120		7268685	7272968	

19

		6	ENSGACG00000017101	*kcnip3b*	7241177	7250465	Kv channel interacting protein 3b, calsenilin
		7	ENSGACG00000017093	*trim69*	7234209	7237347	tripartite motif containing 69
		8	ENSGACG00000017091	*bmp1b*	7223141	7233079	bone morphogenetic protein 1b
		9	ENSGACG00000017126	*nfkbil1*	7290042	7291957	nuclear factor of kappa light polypeptide gene enhancer in B-cells inhibitor-like 1
		10	ENSGACG00000017132	*atp6v0a2a*	7293761	7300431	ATPase, H+ transporting, lysosomal V0 subunit a2a
		11	ENSGACG00000017088	*antxr1b*	7212387	7221007	anthrax toxin receptor 1b
		12	ENSGACG00000017144	*osbp2*	7302479	7311793	oxysterol binding protein 2
J	chrXVII	1	ENSGACG00000007398	*foxj3*	6755976	6790998	forkhead box J3
		2	ENSGACG00000007391		6746613	6750817	
		3	ENSGACG00000007405	*ppcs*	6818104	6820519	phosphopantothenoylcysteine synthetase
		4	ENSGACG00000007417	*utp3*	6819603	6822503	UTP3, small subunit (SSU) processome component, homolog (S. cerevisiae)

20

		5	ENSGACG00000007358	*syn2b*	6695094	6740553	synapsin IIb
		6	ENSGACG00000007429		6824354	6825105	
		7	ENSGACG00000007430		6826808	6829333	
		8	ENSGACG00000007437	*si:dkey-264d12.4*	6830166	6831109	epithelial membrane protein 3
		9	ENSGACG00000007365	*TIMP4*	6726433	6733199	TIMP metallopeptidase inhibitor 4
K	chrVII	1	ENSGACG00000020350	*rtn2b*	19983607	19986866	reticulon 2b
		2	ENSGACG00000020349	*ppm1nb*	19977504	19981930	protein phosphatase, Mg2+/Mn2+ dependent, 1Nb (putative)
		3	ENSGACG00000020348	*kcnk12l*	19973787	19975812	potassium channel, subfamily K, member 12 like
		4	ENSGACG00000020351	*pvrl3b*	19998501	20009445	poliovirus receptor-related 3b
		5	ENSGACG00000020347	*itpkca*	19968056	19972062	inositol-trisphosphate 3-kinase Ca
		6	ENSGACG00000020346	*ccdc61*	19964556	19968807	coiled-coil domain containing 61

21

		7	ENSGACG00000020345		19945027	19963823	
		8	ENSGACG00000020352		20013617	20014501	
		9	ENSGACG00000020353	*ppme1*	20017457	20022592	protein phosphatase methylesterase 1
		10	ENSGACG00000020354	*ucp2*	20026028	20032135	uncoupling protein 2
		11	ENSGACG00000020344	*pls3*	19930922	19943010	plastin 3 (T isoform)
		12	ENSGACG00000020355	*dnajb13*	20031874	20034052	DnaJ (Hsp40) homolog, subfamily B, member 13
		13	ENSGACG00000020356	*rab6a*	20035866	20042328	RAB6A, member RAS oncogene family
L	chrXV	1	ENSGACG00000013078		16218640	16219585	
		2	ENSGACG00000013081	*vrk1*	16187758	16194595	vaccinia related kinase 1
		3	ENSGACG00000013067	*ak7b*	16254902	16265808	adenylate kinase 7b
M	chrXVII	1	ENSGACG00000007138	*atp2b2*	6457686	6506154	ATPase, Ca++ transporting, plasma membrane 2

22

		2	ENSGACG00000007204	*slc6a11b*	6519657	6535112	solute carrier family 6
N	chrXI	1	ENSGACG00000008462	*tubg1*	6189885	6195320	tubulin, gamma 1
		2	ENSGACG00000008473	*si:ch211-18i17.2*	6197763	6222667	pleckstrin homology, MyTH4 and FERM domain containing H3
		3	ENSGACG00000008483	*cntnap1*	6237259	6246630	contactin associated protein 1
		4	ENSGACG00000008492	*ezh1*	6251070	6261579	enhancer of zeste 1 polycomb repressive complex 2 subunit
		5	ENSGACG00000008501	*ramp2*	6265046	6267944	receptor (G protein-coupled) activity modifying protein 2
		6	ENSGACG00000008510		6275474	6275799	
		7	ENSGACG00000008514		6277328	6279125	
		8	ENSGACG00000008517	*c1ql3b*	6309449	6320441	complement component 1, q subcomponent-like 3b
		9	ENSGACG00000008519	*ccdc43*	6382036	6384675	coiled-coil domain containing 43

23

		10	ENSGACG00000008523	*fzd2*	6393118	6394227	frizzled class receptor 2
		11	ENSGACG00000008527	*mylk5*	6409942	6413758	myosin, light chain kinase 5
		12	ENSGACG00000008532	*si:ch73-141c7.1*	6416503	6418554	si:ch73-141c7.1
		13	ENSGACG00000008535	*hsd17b1*	6419439	6421734	hydroxysteroid (17-beta) dehydrogenase 1
		14	ENSGACG00000008544	*zgc:153952*	6439379	6447745	zgc:153952
		15	ENSGACG00000008553	*atp6v0a1a*	6456751	6466815	ATPase, H+ transporting, lysosomal V0 subunit a1a
		16	ENSGACG00000008605	*PTRF*	6468622	6478671	polymerase I and transcript release factor
		17	ENSGACG00000008607	*stat3*	6484785	6492597	signal transducer and activator of transcription 3 (acute-phase response factor)
		18	ENSGACG00000008634	*stat5a*	6519466	6529816	signal transducer and activator of transcription 5a
		19	ENSGACG00000008641	*si:ch211-210g13.5*	6567077	6585055	si:ch211-210g13.5
		20	ENSGACG00000008648	*kcnh4a*	6590904	6602856	potassium voltage-gated channel, subfamily H (eag-related), member 4a

24

O	chrII	1	ENSGACG00000015507	*kif18a*	9100586	9119396	kinesin family member 18A
		2	ENSGACG00000015505		9086923	9092138	
		3	ENSGACG00000015510	*mettl15*	9139122	9161054	methyltransferase like 15
		4	ENSGACG00000015502	*bdnf*	9073816	9074815	ribosomal protein, large P2, like
		5	ENSGACG00000015500	*lin7c*	9068964	9070913	lin-7 homolog C (C. elegans)
		6	ENSGACG00000015499	*rplp2l*	9063720	9064814	ribosomal protein, large P2, like
P	chrI	1	ENSGACG00000009072	*grik4*	8598417	8683492	glutamate receptor, ionotropic, kainate 4
Q	chrIX	1	ENSGACG00000018024	*ugt8*	12576700	12584061	UDP glycosyltransferase 8
		2	ENSGACG00000018022	*ndst3*	12647423	12676796	N-deacetylase/N-sulfotransferase (heparan glucosaminyl) 3
R	chrI	1	ENSGACG00000014605		25581534	25587141	
		2	ENSGACG00000014627	*cbsb*	25574232	25581159	cystathionine-beta-synthase b

25

		3	ENSGACG00000014641		25563835	25564550	
		4	ENSGACG00000014600	*zgc:172122*	25627729	25630447	
		5	ENSGACG00000014598		25633219	25652062	
S	chrXVI	1	ENSGACG00000005749	*WDSUB1*	11001204	11010414	WD repeat, sterile alpha motif and U-box domain containing 1
		2	ENSGACG00000005757	*TANC1*	11011679	11064194	tetratricopeptide repeat, ankyrin repeat and coiled-coil containing 1
		3	ENSGACG00000005734	*BAZ2B*	10980019	10999671	bromodomain adjacent to zinc finger domain 2B

26

Table S5: Characterization of the 42 core SNPs and frequency of the acidic alleles in marine stickleback (marine samples pooled). The SNPs are sorted by decreasing magnitude of B-A differentiation. The colors coding indicates whether at a given SNP, the major allele in the sea coincides with the one typical of the basic (blue) or acidic (red) populations.

Chr	SNPpos	Mean B-A AFD	Acidic allele (based on global FW pool)	Basic allele (based on global FW pool)	Marine minor allele (based on the 30 marine individuals)	Marine major allele (based on the 30 marine individuals)	Marine minor allele count	Marine major allele count	% Acidic allele in marine pop	% Basic allele in marine pop
chrIX	13360688	0.863	G	C	G	C	3	143	0.021	0.979
chrI	879044	0.826	C	A	A	C	25	77	0.755	0.245
chrIV	26641811	0.818	G	A	G	A	3	171	0.017	0.983
chrVII	13825503	0.800	C	T	C	T	46	102	0.311	0.689
chrXX	10619356	0.793	T	C	T	C	6	117	0.049	0.951

27

chrV	3953444	0.791	A	G	A	G	31	119	0.207	0.793
chrIV	28685877	0.791	G	A	G	A	5	141	0.034	0.966
chrIV	12031152	0.787	C	G	C	G	8	143	0.947	0.053
chrXIV	7260519	0.784	A	G	A	G	31	110	0.220	0.780
chrXVII	6782419	0.781	C	A	A	C	1	134	0.993	0.007
chrVII	19986534	0.779	G	A	G	A	67	75	0.472	0.528
chrXV	16209497	0.778	T	A	T	A	11	115	0.087	0.913
chrXVII	6492561	0.776	A	G	G	A	29	127	0.814	0.186
chrXI	6536822	0.769	T	G	T	G	2	152	0.013	0.987
chrII	9113855	0.760	G	A	G	A	36	120	0.231	0.769
chrI	8680374	0.757	G	A	G	A	29	146	0.166	0.834
chrIX	12615477	0.754	T	C	C	T	46	112	0.709	0.291

28

chrI	25584840	0.753	A	T	A	T	10	137	0.068	0.932
chrXVI	11017185	0.752	C	T	C	T	27	83	0.245	0.755
chrX	7862245	0.747	A	G	A	G	31	118	0.208	0.792
chrVII	3036974	0.741	C	A	C	A	65	70	0.481	0.519
chrVII	13908696	0.738	C	G	C	G	43	102	0.297	0.703
chrIV	7849606	0.732	T	G	T	G	23	106	0.178	0.822
chrI	21308259	0.729	C	A	C	A	46	67	0.407	0.593
chrXV	881351	0.728	A	G	A	G	5	136	0.035	0.965
chrIV	8523376	0.727	A	C	A	C	40	87	0.315	0.685
chrVII	14277048	0.725	C	A	C	A	11	157	0.065	0.935
chrIX	13102705	0.720	C	A	A	C	1	189	0.995	0.005
chrIV	7010171	0.718	C	T	C	T	70	80	0.467	0.533

29

chrVII	5378300	0.718	A	G	A	G	18	40	0.310	0.690
chrX	11336552	0.717	A	G	G	A	6	124	0.954	0.046
chrIII	5077315	0.716	T	G	T	G	25	61	0.291	0.709
chrVII	22905391	0.713	A	C	C	A	62	82	0.569	0.431
chrX	9814424	0.712	A	G	G	A	1	108	0.991	0.009
chrXVII	1404567	0.712	T	A	T	A	10	90	0.100	0.900
chrXX	13066370	0.709	C	A	C	A	48	110	0.304	0.696
chrXX	9770450	0.709	C	T	C	G	5	112	0.045	0.000
chrVII	12965594	0.706	C	T	C	T	29	45	0.392	0.608
chrXX	9376877	0.704	T	G	T	G	47	90	0.343	0.657
chrV	8832730	0.700	G	A	G	A	45	95	0.321	0.679
chrIV	10565725	0.700	A	C	A	C	33	125	0.209	0.791

30

chrXVI	9076916	0.700	A	G	A	G	50	96	0.342	0.658

31

Figures

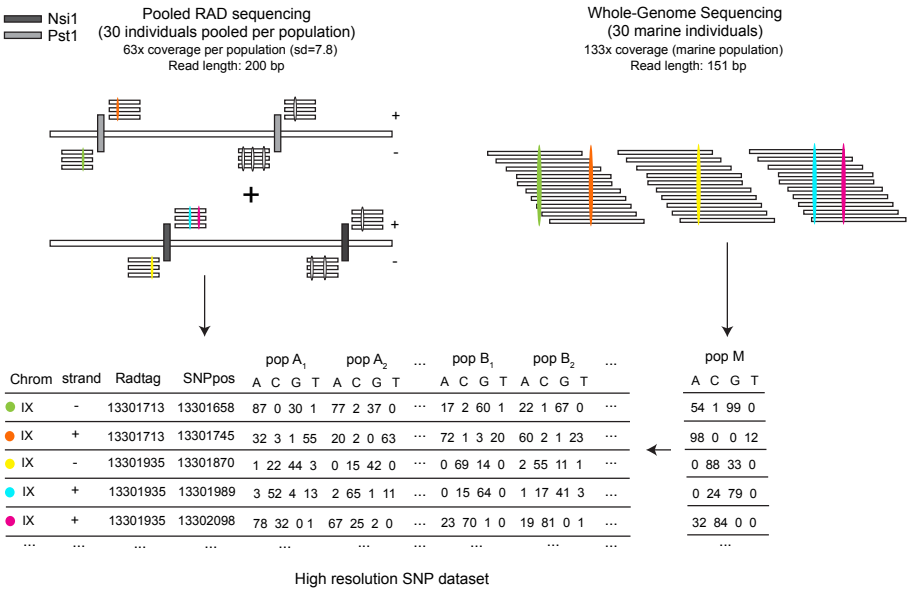


Figure S1: Schematic description of the SNP generation protocol based on pooled RAD and whole-genome sequencing. For the basic and acidic populations, we modified the classical RAD protocol (Baird et al 2008) by performing a parallel digestion of the basic and acidic samples (~ 30 individuals pooled per population) by two restriction enzymes (Nsi1 and Pst1) (top left). For the 30 total marine individuals, we performed whole-genome sequencing (top right). SNPs are visualized as colored ovals. After appropriate filtering steps, a high-resolution SNP dataset was then generated by performing allele counts for each population at each base position (bottom left). Marine allele counts were performed only at the SNPs ascertained in the freshwater samples and added to the SNP matrix (bottom right).

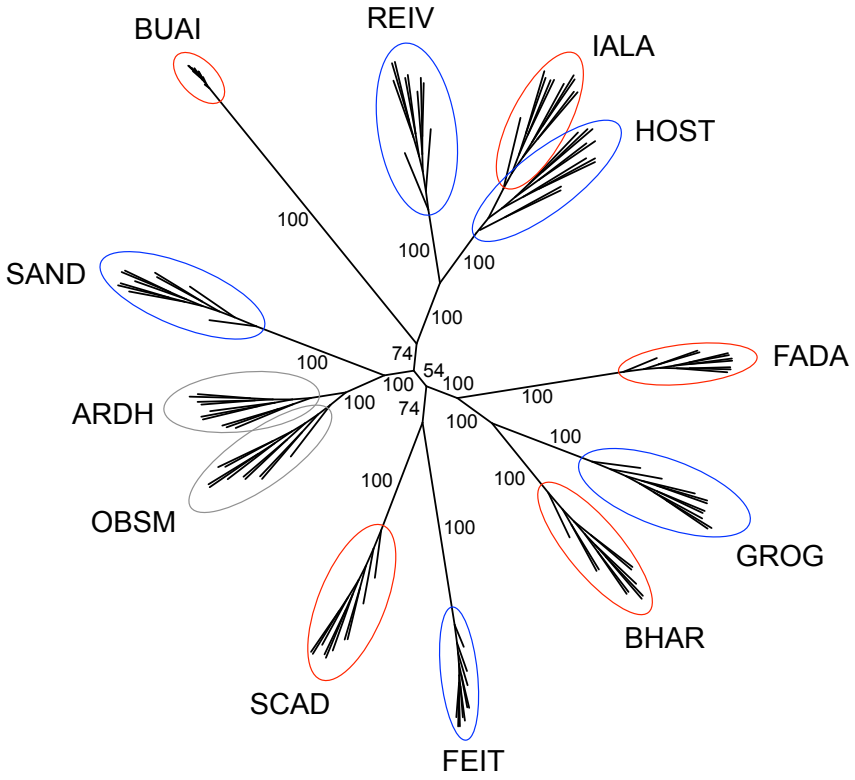
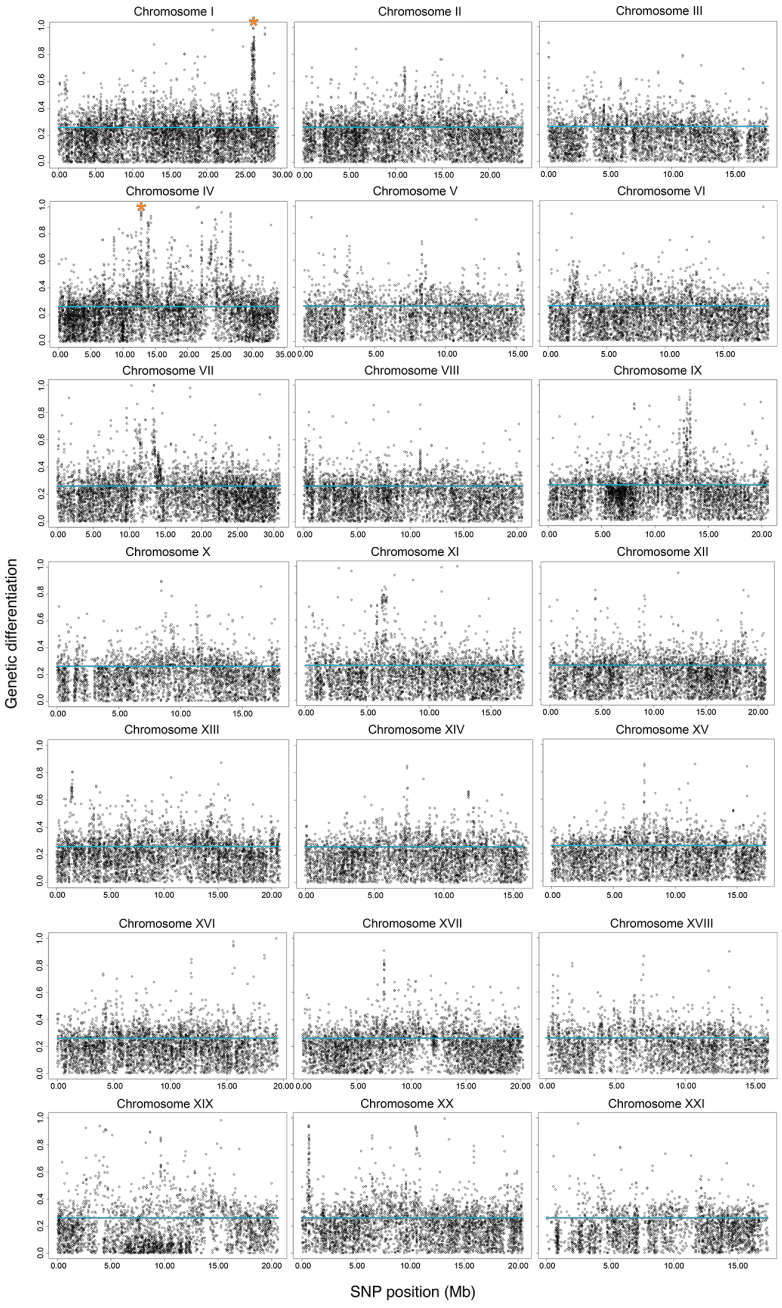


Figure S2: Unrooted nuclear phylogeny based on 15,058 SNPs, using the full ten synthetic individuals generated for each population (instead of a single one, as in Fig. 2A). Color coding is by habitat, as in Fig. 1.

Figure S3


A
Chromosome I
Chromosome II
Chromosome III
Chromosome IV
Chromosome V
Chromosome VI
Chromosome VII
Chromosome VIII
Chromosome IX
Chromosome X
Chromosome XI
Chromosome XII
Chromosome XIII
Chromosome XIV
Chromosome XV
Chromosome XVI
Chromosome XVII
Chromosome XVIII
Chromosome XIX
Chromosome XX
Chromosome XXI
Genetic differentiation
SNP position (Mb)

34

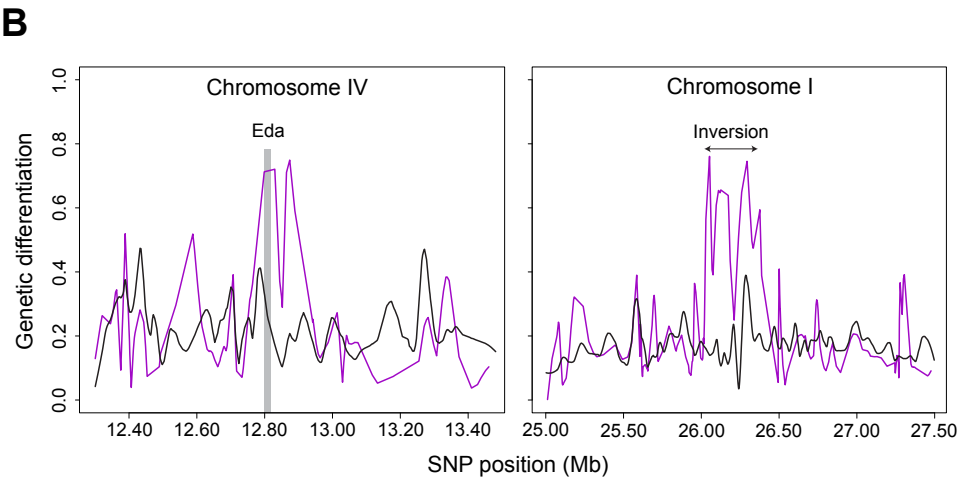


Figure S3: (A) Profiles of marine-freshwater genetic differentiation, quantified as absolute allele frequency difference, across the 21 treespine stickleback chromosomes. The blue line indicates the median (0.26) across all genome-wide SNPs. For this analysis, all basic and acidic populations were combined to a global freshwater pool and compared to the marine population (i.e., the OBSM and ARDH samples pooled). SNP detection followed a strategy differing from the one underlying the SNP data set used for the main investigation of basic-acidic differentiation: at each base position covered by the RAD tags of the freshwater fish, we considered the full marine nucleotide coverage (average: 133x), and a nucleotide sample of exactly the same size drawn at random from the freshwater pool. A variable position then qualified as SNP when these two samples combined exhibited a MAF of at least 0.05, and when read coverage was within 50-240 for the marine pool and within 200-2800 for the full freshwater pool. For SNPs satisfying these criteria, we then performed base counts for each population (1 marine, 10 freshwater). These data were saved in a SNP matrix (available on Dryad) and used to calculate overall marine versus freshwater allele frequency differences at all SNPs. Dark orange

asterisks indicate the position of the chromosome I inversion and the Eda locus (chromosome IV), regions well-known to be under divergent selection between marine and freshwater stickleback (e.g., Hohenlohe et al. 2010; Jones et al. 2012; Roesti et al. 2014; Terekhanova et al. 2014; Nelson & Cresko 2018). (B) Patterns of genetic differentiation around the same two classical loci of marine-freshwater differentiation, based on the SNPs ascertained using the freshwater populations only (i.e., as in our main analyses). The purple and black lines here represent marine-freshwater and basic-acidic (B-A) differentiation (mean AFD across all corresponding population comparisons). Note that B-A differentiation is low at both loci, consistent with the sharing of haplotypes *universally* favorable in freshwater (that is, favorable in both basic and acidic lakes). The profiles are smoothed to reduce complexity; the magnitude of marine-freshwater differentiation at individual SNPs is even higher.

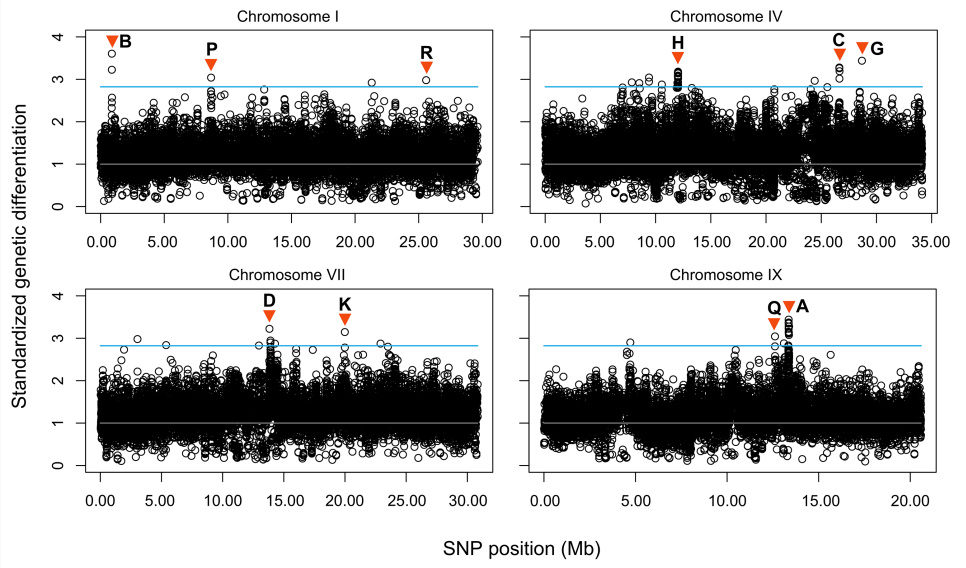


Figure S4: Check of the robustness of identifying genomic regions of highly parallel basic-acidic differentiation (top core SNPs) by integrating AFD data from multiple population comparisons without taking differences among comparisons in their overall level of differentiation into account (the outcome of this type of data integration chosen for our study is hereafter called 'AFD_{RAW}'). For this, we repeated the integration of AFD data across the B-A comparisons by first standardizing all AFD values from a given population comparison by the genome-wide median AFD value for that comparison (yielding 'AFD_{STAND}'). As a critical check for the consistency between the non-standardized and standardized identification of the core SNPs, we then retrieved the top 42 SNPs based on AFD_{STAND} (corresponding to a threshold of 2.8241, indicated by the blue line above) and determined the degree of overlap with the 42 SNPs identified based on AFD_{RAW} (i.e., the normal core SNPs in the main paper). The congruence between these two approaches was very high: 36 (86%) of

the 42 top SNPs identified using the AFD_{STAND} approach proved identical at the precise base pair level with our core SNPs. The similarity between the two approaches is visualized above for the four chromosomes harboring the highest number of top core SNPs. Here the dots represent the average B-A differentiation across the multiple comparisons at each SNP, as obtained after standardizing each comparison by its genome-wide median (hence the Y-axis scale no longer ranges from zero to one, contrary to AFD_{RAW} in Fig. 3A). Dark orange triangles indicate the position of the core SNPs on these chromosomes, as based on the AFD_{RAW} approach. These SNPs also emerge as the regions of strongest differentiation on each chromosome when using the AFD_{STAND} method. The high consistency between the two approaches to integrating differentiation data from multiple population comparisons justifies using the mathematically simpler one (i.e., no standardization). A further reason why we base our identification of top core SNPs on AFD_{RAW} is that the core SNPs of these genome regions proved completely fixed for alternative alleles (i.e., AFD = 1) in some basic-acidic population comparisons. Since AFD cannot increase beyond one even when the overall level of differentiation continues to increase, standardization by the latter may lead to the underestimation of genetic differentiation. Given that the overall level of differentiation was reasonable similar among all B-A comparisons anyway (Table S2), the non-standardized approach appeared superior to us.

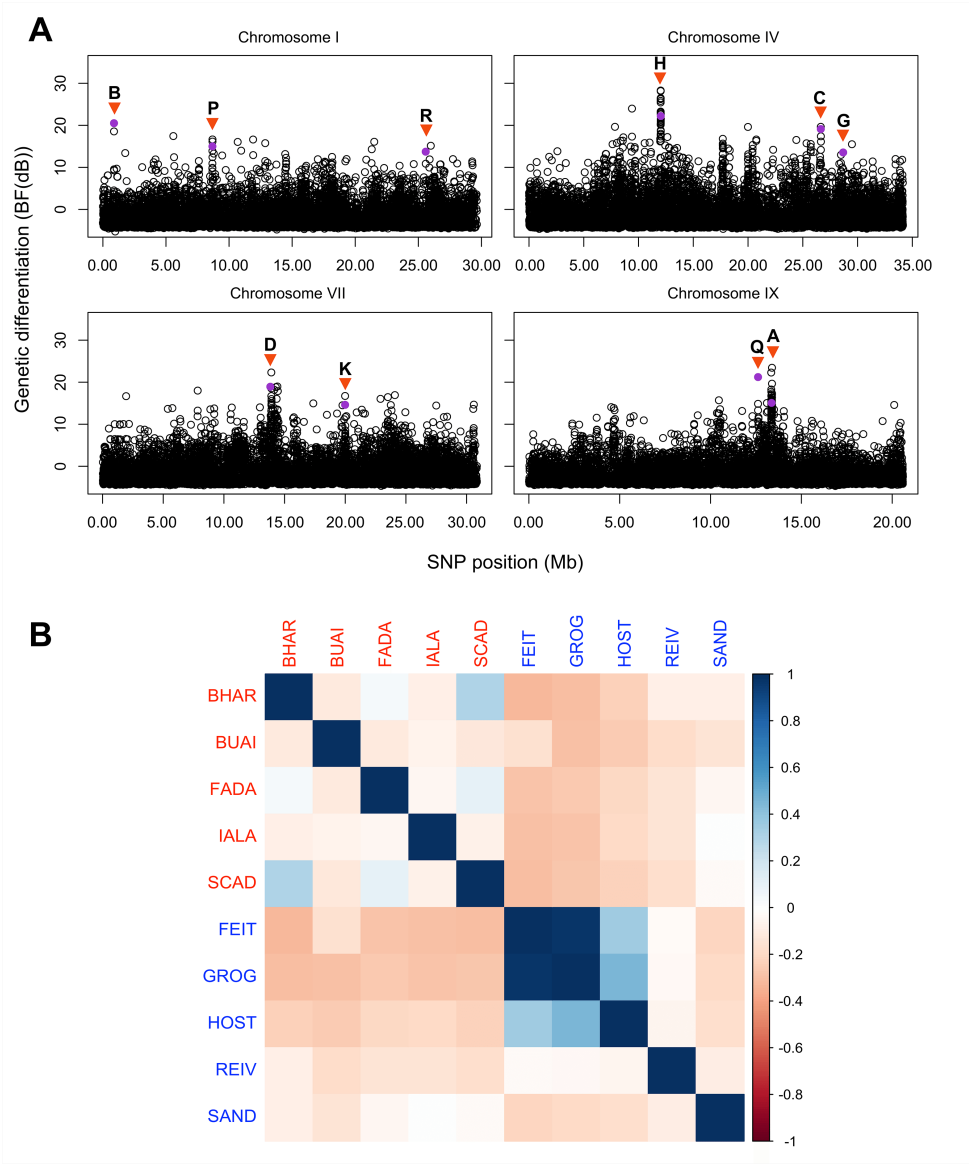


Figure S5: BayPass analysis. (A) Check of the robustness of identifying genomic regions of highly parallel basic-acidic differentiation (i.e., core SNP regions) by integrating AFD data from multiple population comparisons (our method presented in the paper) against an ecotype-related outlier SNP scan using BayPass (Gautier

2015). For the BayPass analysis, we used the same SNP data set as for our method, a binary scoring for the ecotype covariate (1 = basic, 2 = acidic), and the same parameters as described in Leblois et al. 2017. As a critical check for the consistency between the two approaches, we focused on the variable BF(dB) from the BayPass output expressing for each SNP the strength of association to basic versus acidic ecotype. Based on this variable, we retrieved the top 1% of the SNPs and determined visually what proportion of our 42 core SNP regions coincided with regions containing one or multiple of these BayPass 'outliers'. This check revealed a high congruence between the methods: 36 (86%) out of our 42 total core SNP regions also emerged unambiguously as BayPass outlier regions. In (A), this congruence is visualized for the same four chromosomes as in Fig. S4. Here, dark orange triangles indicate a subsample of our 19 top core SNP regions (i.e., core SNP showing AFD > 0.75 in the combined B-A comparison), with the precise core SNPs shown as purple dots. The consistency between the methods clearly confirms the robustness of our method. (B) Correlation matrix based on scaled population allele frequencies covariances estimated by BayPass. The color shade expresses the magnitude of positive or negative correlation for a given population pair. This matrix generally reveals weak among-population correlations in allele frequencies, as expected from the independent evolution of the lake populations indicated by our other analyses (phylogenies, ordination). A potential lack of independence is suggested only for the FEIT and GROG basic population pair. This appears plausible, given that the outlet of FEIT could not be determined with confidence (Fig. 1A).

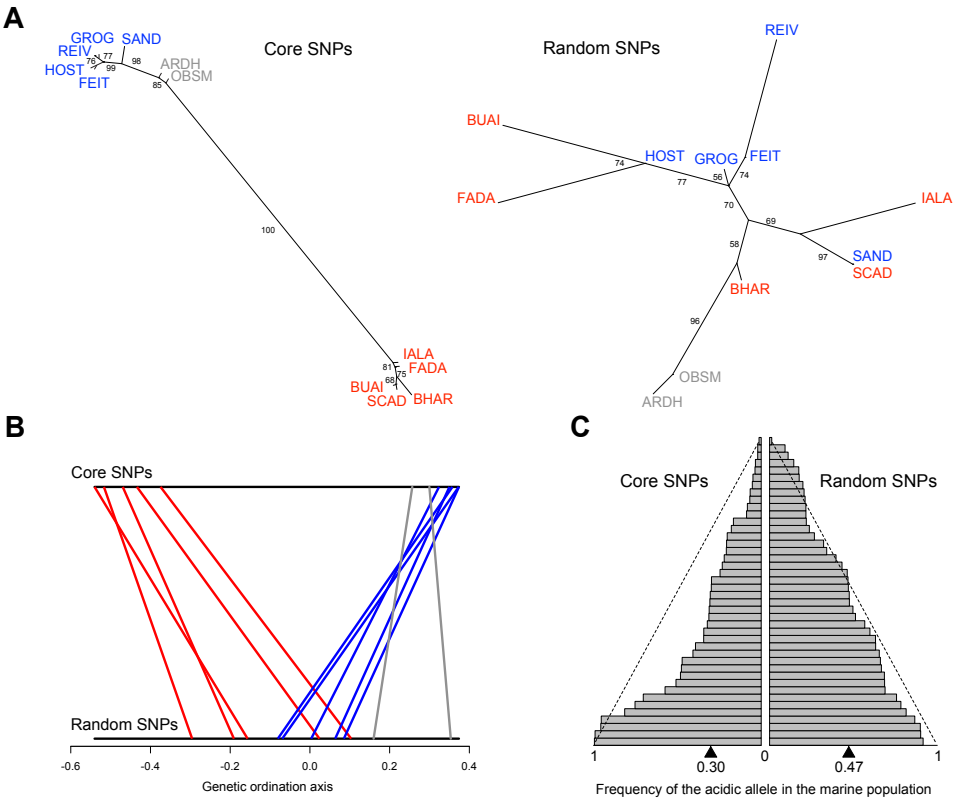


Figure S6: Replication of the analyses presented in Fig. 4 based on a new set of random SNPs. These were chosen at random among all the SNPs displaying an AFD inferior to 0.5 in the integrated B-A comparison. We here thus controlled much less effectively for the selective neutrality of the random SNPs (recall that the random SNPs used in the paper were required to fall within a very narrow AFD window around the genome-wide median). Apart from the different set of random SNPs, all analytical conventions and graphing styles correspond to those underlying Fig. 4. Note that using a different set of random SNPs leads to similar results supporting the same conclusion, even when enforcing the selective neutrality of these SNPs less strictly.

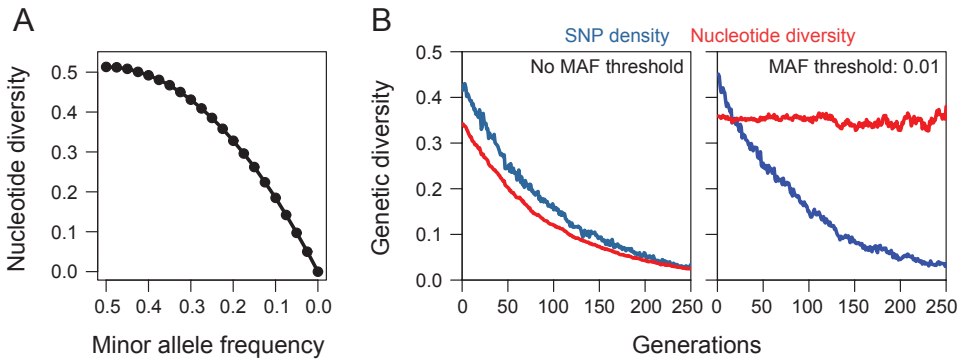


Figure S7: Exploration of the sensitivity and robustness of SNP density, the metric of within-population genetic diversity employed in our study, in comparison to nucleotide diversity (π; Nei & Li 1979). (A) Shows nucleotide diversity, computed as the fraction of nucleotide mismatches among all possible pairwise nucleotide permutations, along the continuum of decreasing genetic variation as defined by the frequency of the minor allele (MAF) among 40 total nucleotides. The left end of the X-axis represents two alleles in perfectly balanced proportion (20 vs. 20), while the right end corresponds to the fixation for one allele (40 vs. 0). This numerical analysis reveals a non-linear response of nucleotide diversity to the loss of genetic variation at a polymorphism: a given allele frequency reduction causes a relatively weak change in nucleotide diversity in the MAF range representing alleles in relatively balanced proportion, whereas an allele frequency reduction of the same magnitude drives a strong change in nucleotide diversity in the MAF range in which one allele is rare. In (B), SNP density – a genetic diversity metric derived directly from the MAF, and nucleotide diversity were applied to a simulated population to examine how these metrics respond to a reduction in genetic diversity across numerous loci. We here simulated a population of 100 haploid individuals and 1000 unlinked bi-allelic SNPs. At each SNP, genotypes were initially drawn at random from a uniform distribution

(for an empirical justification see Fig. 4C). The population then experienced a loss of diversity over 250 generations by drift, achieved by re-sampling each SNP with replacement to the original population size. In each generation, a subsample of 40 nucleotides (similar to the minimum coverage threshold used in our empirical analyses) was drawn at each SNP. Based on these subsamples, nucleotide diversity was calculated as described above and averaged over all SNPs. SNP density was calculated as the proportion of SNPs for which the subsample satisfied a MAF threshold of 0.3 (the same threshold as in our empirical analyses of genetic diversity). This algorithm was carried out in two modes: either by accepting *all* SNPs for genetic diversity calculation (visualized in the left panel), or by first filtering the subsample at each SNP according to a mild MAF threshold of 0.01, and calculating the two diversity metrics only based on those SNPs satisfying this threshold (shown in the right panel). With a sample size of 40 nucleotides, this latter MAF threshold eliminated all monomorphic SNPs plus the singletons.

The left panel of (B), involving no low-MAF filter, shows that as diversity declines (i.e., the SNPs move stochastically toward monomorphism), SNP density and nucleotide diversity are tightly correlated. Consistent with the reduced sensitivity of nucleotide diversity to allele frequency shifts in the high-MAF range identified in (A), however, the decline in nucleotide diversity is slightly less steep than the decline in SNP density. At least for SNPs showing allele frequencies broadly consistent with a uniform distribution, SNP density thus captures the loss of diversity more sensitively than nucleotide diversity.

The right panel of (B) further reveals a dramatic influence of low-MAF filtering on nucleotide diversity but not SNP density: excluding monomorphic SNPs and

singletons renders nucleotide diversity almost completely insensitive to diversity reduction. Although the mild MAF filter (0.05) applied to the global pool of all our freshwater populations to exclude sequencing error is unlikely to eliminate low-diversity sites within the populations as radically as the low-MAF filter in this second simulation mode, this simulation nevertheless makes clear that MAF thresholds can affect the estimation of genetic diversity by nucleotide diversity substantially. The reason is that such thresholds alter both total SNP number and the relative fraction of those SNPs for which nucleotide diversity exhibits the highest sensitivity (i.e., the low-MAF range, see A). By contrast, SNP density is not materially influenced by MAF filtering.

Overall, we conclude that SNP density, the metric of genetic diversity adopted in our work, not only captures diversity loss more sensitively than nucleotide diversity, it also represents a diversity metric highly robust to MAF filtering. Clearly, the use of SNP density in our analytical context is well motivated.

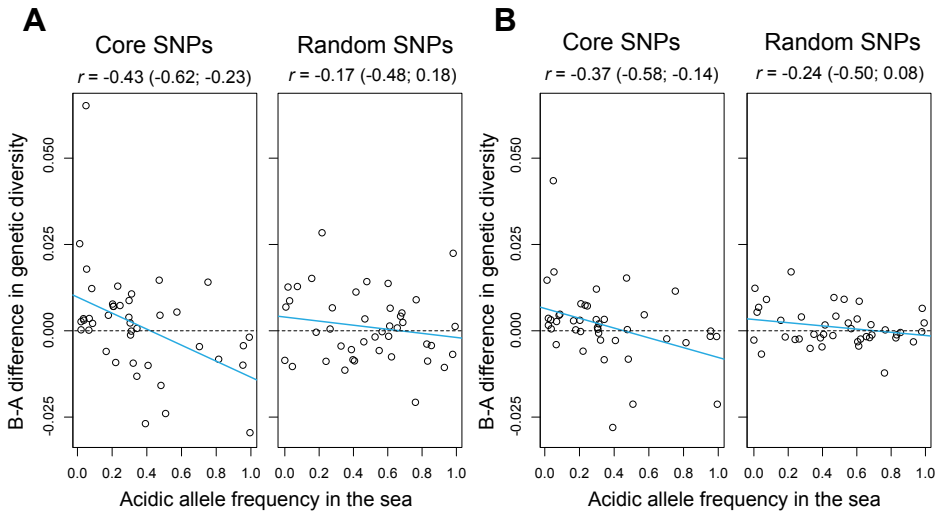


Figure S8: Robustness of the selective sweep analysis. In (A), this analysis was repeated by considering SNP density across a narrower chromosome window (20 kb as opposed to 40 kb) around the focal SNPs (core and random). In (B), we performed the selective sweep analysis by applying a different MAF threshold for determining the number of high-MAF SNPs (0.2 as opposed to 0.3). Further MAF thresholds examined included 0.15 and 0.25, producing similar results, although for theoretical reasons mentioned in the paper, high MAF thresholds should reveal selective sweeps most reliably. All other analytical conventions and graphing styles follow those underlying Fig. 6. Collectively, these supplementary analyses confirm a strong relationship between genetic diversity and allele frequencies in the sea for the core SNPs only, consistent with our conclusion of selective sweeps drawn in the paper.

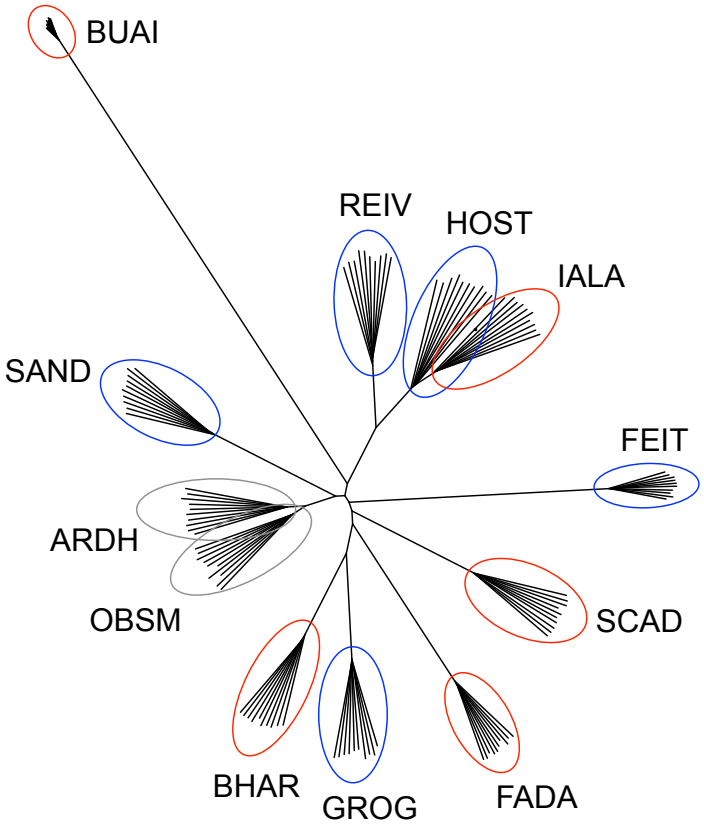


Figure S9: Unrooted neighbor-joining nuclear phylogeny based on 15,058 SNPs, using the full ten synthetic individuals generated for each population. Color coding is by habitat, as in Fig. 1.

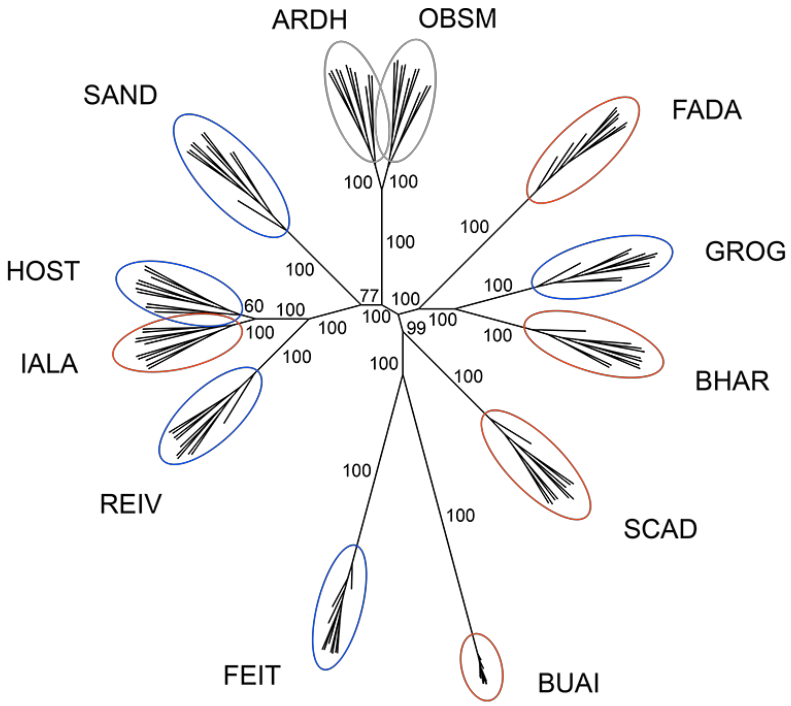


Figure S10: Unrooted nuclear phylogeny based on 68,245 SNPs, using the full ten synthetic individuals generated for each population. Color coding is by habitat, as in Fig. 1.

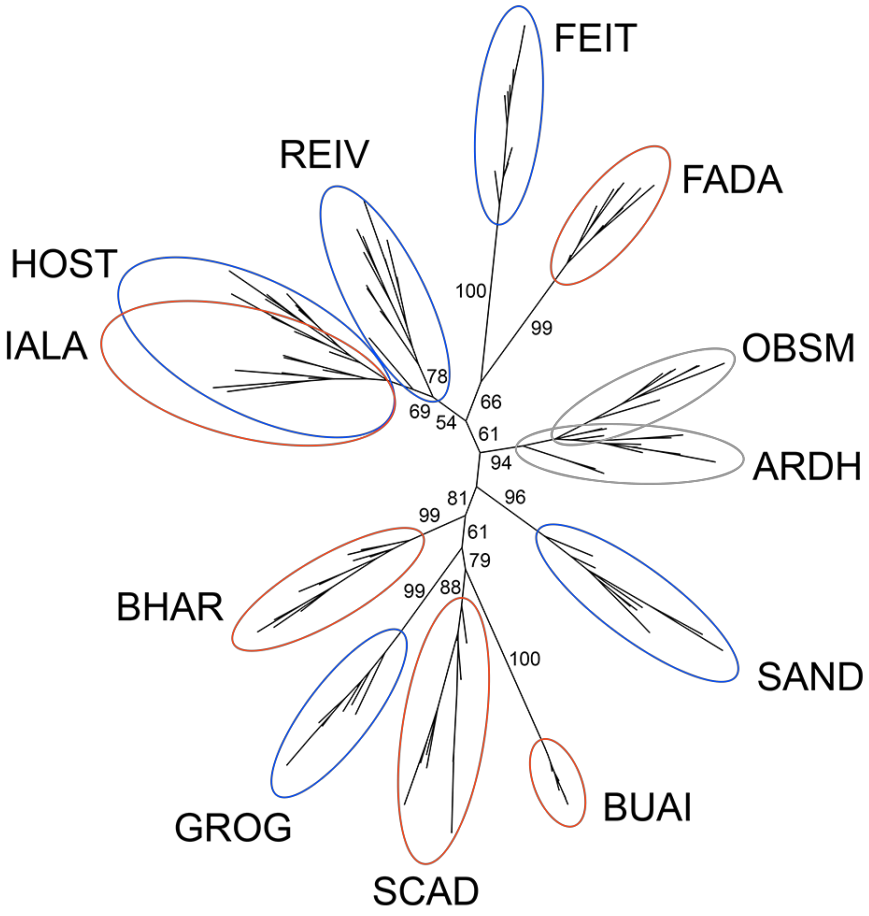
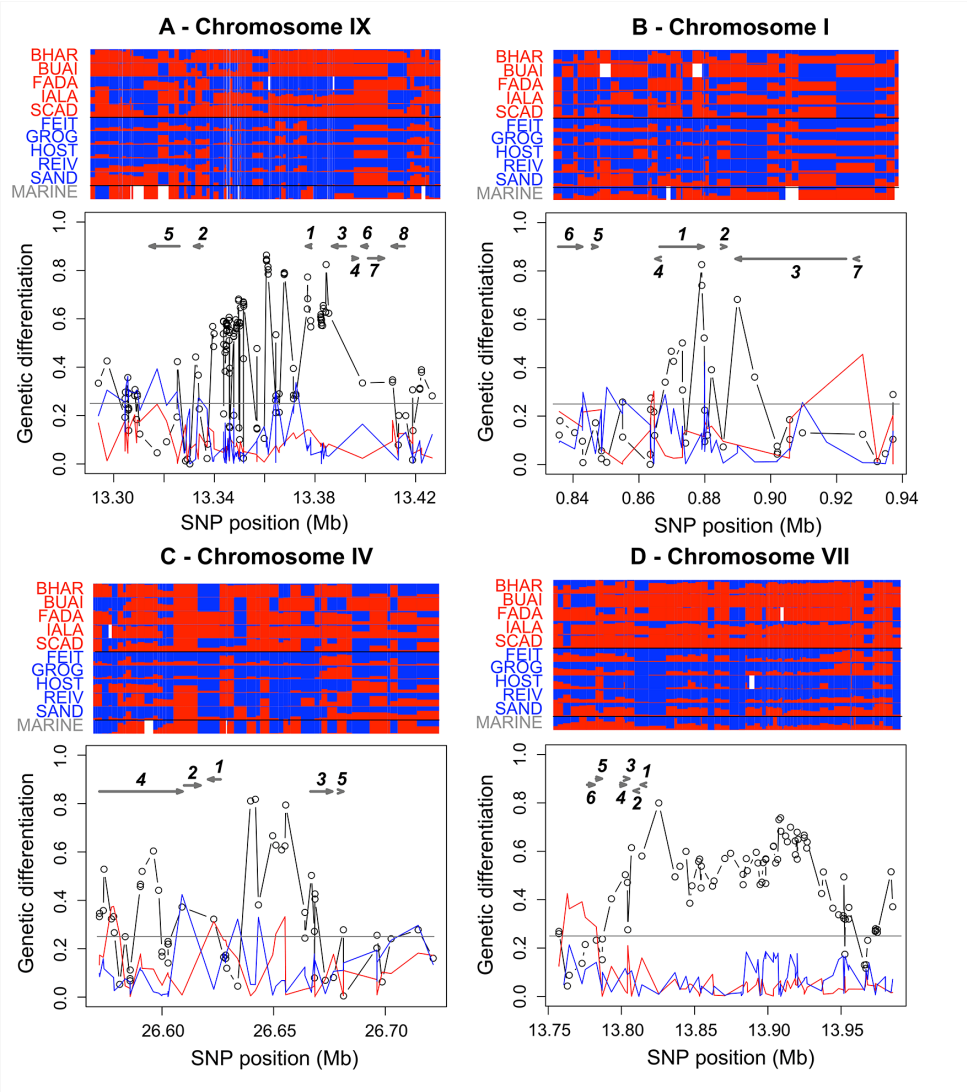
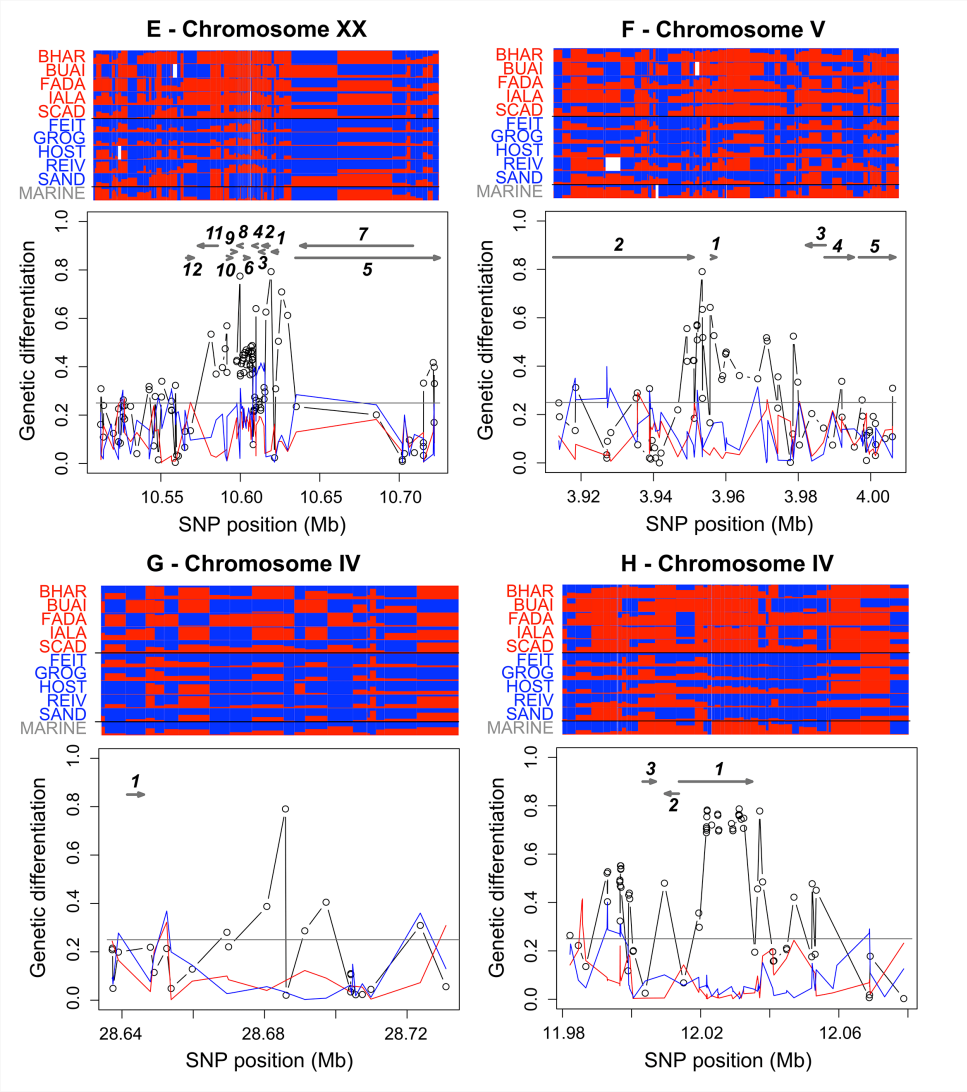


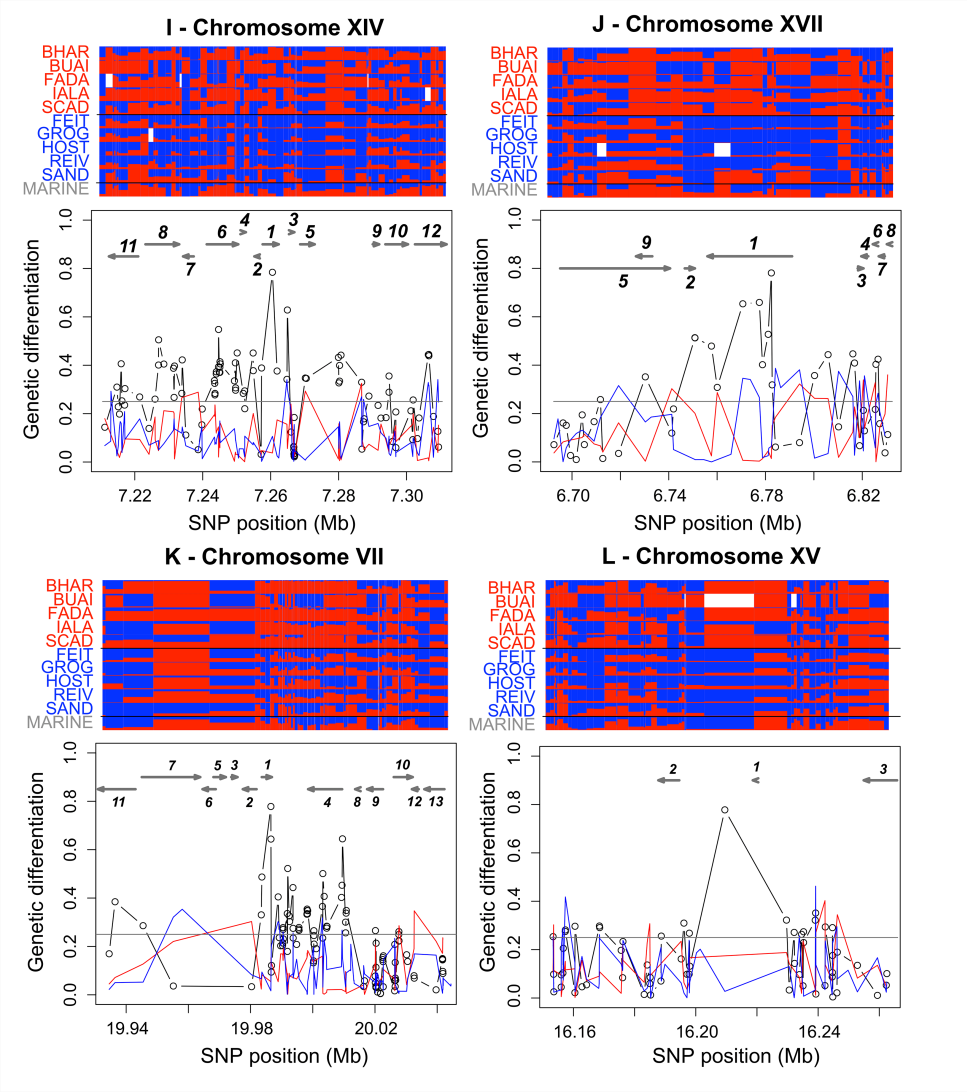
Figure S11: Unrooted nuclear phylogeny based on just 227 SNPs spaced by at least 1 Mb, with ten synthetic individuals generated for each population. Color coding is by habitat, as in Fig. 1. Note that despite this low number of markers, the populations are generally still monophyletic, and the position of basic and acidic populations across the tree remains random, consistent with the independent evolution of the freshwater populations.

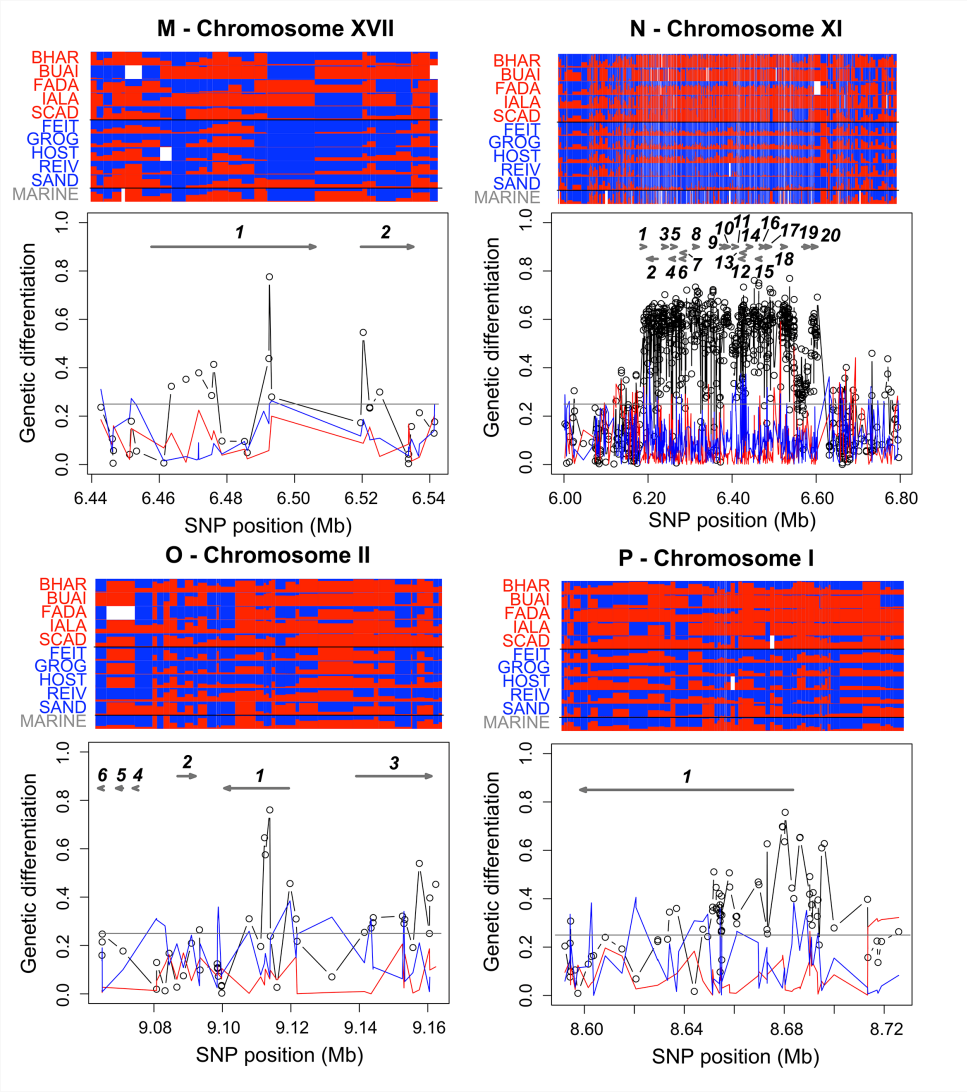
Figure S12


A - Chromosome IX
B - Chromosome I
C - Chromosome IV
D - Chromosome VII
BHAR
BUAI
FADA
IALA
SCAD
FEIT
GROG
HOST
REIV
SAND
MARINE
Genetic differentiation
SNP position (Mb)
13.30
13.34
13.38
13.42
0.84
0.86
0.88
0.90
0.92
0.94
26.60
26.65
26.70
13.75
13.80
13.85
13.90
13.95


E - Chromosome XX
F - Chromosome V
G - Chromosome IV
H - Chromosome IV
BHAR
BUAI
FADA
IALA
SCAD
FEIT
GROG
HOST
REIV
SAND
MARINE
Genetic differentiation
SNP position (Mb)
11 9 8 4 2 1
12 10 6 3
7
5
2
1
3
4
5
10.55
10.60
10.65
10.70
3.92
3.94
3.96
3.98
4.00
28.64
28.68
28.72
11.98
12.02
12.06
0.0
0.2
0.4
0.6
0.8
1.0

50


I - Chromosome XIV
J - Chromosome XVII
K - Chromosome VII
L - Chromosome XV
BHAR
BUAI
FADA
IALA
SCAD
FEIT
GROG
HOST
REIV
SAND
MARINE
Genetic differentiation
SNP position (Mb)


M - Chromosome XVII
N - Chromosome XI
O - Chromosome II
P - Chromosome I
BHAR
BUAI
FADA
IALA
SCAD
FEIT
GROG
HOST
REIV
SAND
MARINE
Genetic differentiation
SNP position (Mb)

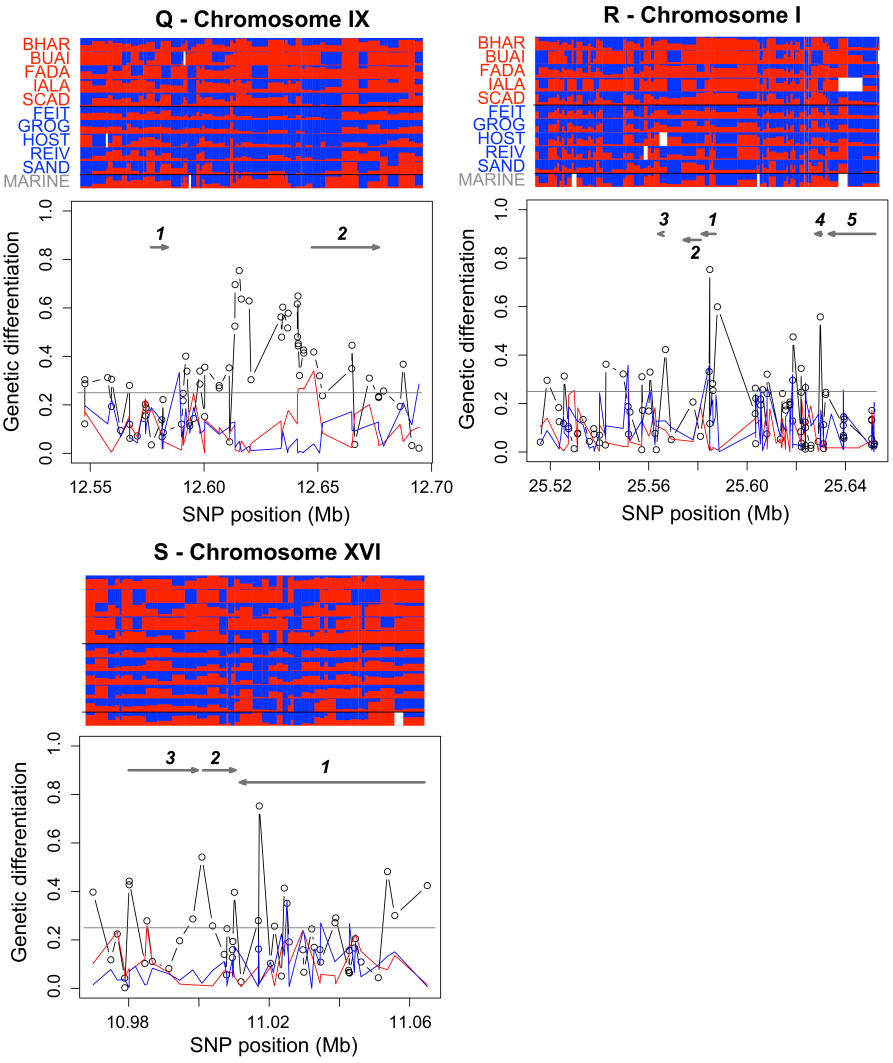


Figure S12: Description of the 19 top core SNPs (mean AFD > 0.75 across the integrated B-A comparison) representing the genomic regions showing the strongest and most consistent basic-acidic differentiation. Regions are ordered by decreasing AFD at the core SNP and are labeled from A to S, consistent with the labeling used

in Table S4. Presentation style follows Fig. 3A, except that genes are numbered to link them to their characterization provided in Table S4.

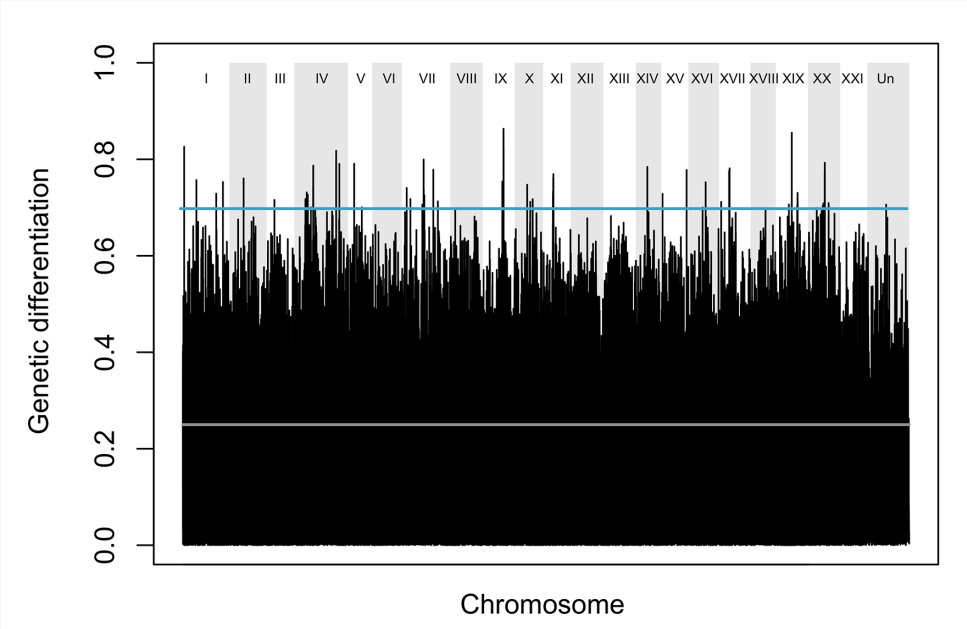


Figure S13: Genome-wide basic-acidic differentiation, as obtained by integrating SNP-specific AFD values across all B-A comparisons. The gray line represents genome-wide median differentiation (0.25), the blue line represents the threshold (0.70) used to identify the core SNPs considered genomic regions of strong and consistent B-A differentiation. Gray and white backgrounds separate the chromosomes. The chromosome 'Un' represents a concatenation of scaffolds not physically anchored to the other chromosomes. The region of high differentiation on chromosome XIX was not considered for further analysis, as this chromosome is the sex chromosome in threespine stickleback.

Figure S14

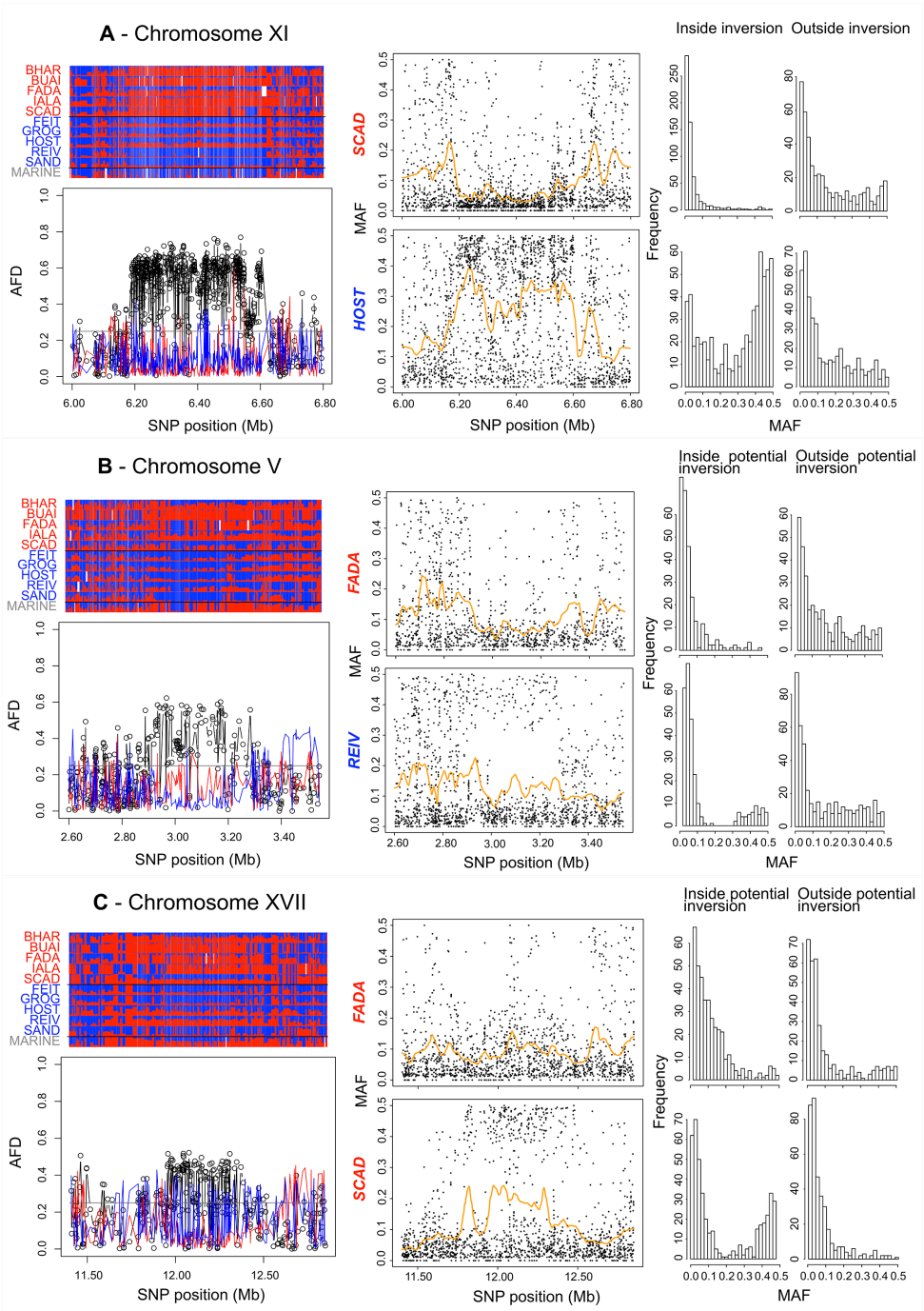

A - Chromosome XI
Inside inversion
Outside inversion
BHAR
BUAI
FADA
IALA
SCAD
FEIT
GROG
HOST
REIV
SAND
MARINE
AFD
SCAD
HOST
MAF
Frequency
SNP position (Mb)
MAF
B - Chromosome V
Inside potential inversion
Outside potential inversion
FADA
REIV
C - Chromosome XVII

Figure S14: Chromosomal inversions produce characteristic patterns in population differentiation and in the frequency of the minor allele, illustrated above for the known inversion on chromosome XI (A), and for two novel potential inversions on the chromosomes V and XVII (B and C). For each (potential) inversion, the bottom panels in the left column present mean genetic differentiation (AFD) profiles for the integrated basic-acidic (black), basic-basic (blue) and acidic-acidic (red) population comparisons, as in Fig. 3A. The dots represent individual SNPs, and the horizontal gray lines indicate genome-wide median differentiation for the integrated B-A comparisons. The top left panels visualize the relative frequencies of the SNPs alleles in each population at all SNPs underlying the differentiation profiles, again following Fig. 3A. The middle column presents the minor allele frequency (MAF) at each SNP position across a chromosome segment around the (potential) inversion. The top panels show MAF for a population (nearly) monomorphic for one specific inversion type, whereas the bottom panels show MAF for a population in which both inversion types occur at relatively balanced frequencies. The right column summarizes the frequency distribution of the MAF across the chromosome segments visualized in the middle column, separately so for the SNPs located inside (left) and outside (right) of the (candidate) inversions (assumed boundaries, from top to bottom: 6.2-6.6 Mb; 2.9-3.3 Mb; 11.9-12.4 Mb). Note that the presence of both (potential) inversion types at a balanced frequency generates a bimodal MAF distribution.

References

Bell, M. F., S. (1994). *The evolutionary biology of the threespine stickleback*, Oxford University Press.

Bierne, N., Gagnaire, P. A. and David, P. (2013). The geography of introgression in a patchy environment and the thorn in the side of ecological speciation. *Curr Zool* 59: 72-86.

Catchen, J., Bassham, S., Wilson, T., Currey, M., O'Brien, C., Yeates, Q. *et al.* (2013). The population structure and recent colonization history of Oregon threespine stickleback determined using restriction-site associated DNA-sequencing. *Mol Ecol* 22: 2864-2883.

Gautier, M. (2015). Genome-wide scans for adaptive differentiation and association analysis with population-specific covariables. *Genetics* 201: 1555-1579

Giles, N. (1983). The possible role of environmental calcium levels during the evolution of phenotypic diversity in Outer Hebridean populations of the three-spined stickleback, *Gasterosteus aculeatus*. *J Zool* 199: 535-544.

Hohenlohe, P. A., Bassham, S., Etter, P. D., Stiffler, N., Johnson, E. A. and Cresko, W. A. (2010). Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLoS Genet* 6: e1000862.

Jones, F. C., Chan, Y. F., Schmutz, J., Grimwood, J., Brady, S. D., Southwick, A. M. *et al.* (2012a). A genome-wide SNP genotyping array reveals patterns of global and repeated species-pair divergence in sticklebacks. *Curr Biol* 22: 83-90.

Jones, F. C., Grabherr, M. G., Chan, Y. F., Russell, P., Mauceli, E., Johnson, J. *et al.* (2012b). The genomic basis of adaptive evolution in threespine sticklebacks. *Nature* 484: 55-61.

Leblois, R., Gautier, M., Rohfritsch, A., Foucaud, J., Burban, C., Galan, M. *et al.* (2018). Deciphering the demographic history of allochronic differentiation in the pine processionary moth Thaumetopoea pityocampa. *Mol Ecol* 27: 264-278

Magalhaes, I. S., Agostino, D. D., Hohenlohe, P. A. and Maccoll, A. D. C. (2016). The ecology of an adaptive radiation of three-spined stickleback from North Uist, Scotland. *Mol Ecol* 25(17): 4319-4336.

Nei, M. (1973). Analysis of gene diversity in subdivided populations. *Proc Natl Acad Sci USA* 70: 3321-3323.

Nei, M., Li, W-H. (1979). Mathematical model for studying genetic variation in terms of restriction endocleases. *Proc Natl Acad Sci USA* 76: 5269-5273.

Nelson, T. C. and Cresko, W. A. (2018). Ancient genomic variation underlies repeated ecological adaptation in young stickleback populations. *Evol Lett* 2: 9-21.

Roesti, M., Gavrilets, S., Hendry, A. P., Salzburger, W. and Berner, D. (2014). The genomic signature of parallel adaptation from shared genetic variation. *Mol Ecol* 23: 3944-3956.

Roesti, M., Kueng, B., Moser, D. and Berner, D. (2015). The genomics of ecological vicariance in threespine stickleback fish. *Nat Commun* 6 : 8767.

Terekhanova, N.V., Logacheva, M.D., Penin, A.A., Neretina, T.V., Barmintseva A.E., Bazykin, G.A. *et al.* (2014). Fast evolution from precast bricks: genomics of young freshwater populations of threespine stickleback *Gasterosteus aculeatus*. *PLoS Genet* 10: e1004696-e1004696.

Waterston, A. H., A., Campbell, R. and Maitland, P. (1979). Inland waters of the Outer Hebrides. *P Roy Soc Edinb B* 77: 329-351.

Chapter 2

The maintenance of standing genetic variation - Gene flow vs. selective neutrality in Atlantic stickleback fish

Haenel et al. 2022, Molecular Ecology

Received: 15 July 2021 | Revised: 20 October 2021 | Accepted: 2 November 2021

DOI: 10.1111/mec.16269

ORIGINAL ARTICLE

MOLECULAR ECOLOGY WILEY

The maintenance of standing genetic variation: Gene flow vs. selective neutrality in Atlantic stickleback fish

Quiterie Haenel[1] | Laurent Guerard[2] | Andrew D. C. MacColl[3] | Daniel Berner[1]

[1]Zoology, Department of Environmental Sciences, University of Basel, Basel, Switzerland

[2]Imaging Core Facility, Biozentrum, University of Basel, Basel, Switzerland

[3]School of Life Sciences, University of Nottingham, Nottingham, UK

Correspondence
Quiterie Haenel and Daniel Berner, Zoology, Department of Environmental Sciences, University of Basel, Basel, Switzerland.
Email: quiterie.haenel@unibas.ch and daniel.berner@unibas.ch

Funding information
Schweizerischer Nationalfonds zur Förderung der Wissenschaftlichen Forschung, Grant/Award Number: 31003A_165826

Abstract

Adaptation to derived habitats often occurs from standing genetic variation. The maintenance within ancestral populations of genetic variants favourable in derived habitats is commonly ascribed to long-term antagonism between purifying selection and gene flow resulting from hybridization across habitats. A largely unexplored alternative idea based on quantitative genetic models of polygenic adaptation is that variants favoured in derived habitats are neutral in ancestral populations when their frequency is relatively low. To explore the latter, we first identify genetic variants important to the adaptation of threespine stickleback fish (*Gasterosteus aculeatus*) to a rare derived habitat—nutrient-depleted acidic lakes—based on whole-genome sequence data. Sequencing marine stickleback from six locations across the Atlantic Ocean then allows us to infer that the frequency of these derived variants in the ancestral habitat is unrelated to the likely opportunity for gene flow of these variants from acidic-adapted populations. This result is consistent with the selective neutrality of derived variants within the ancestor. Our study thus supports an underappreciated explanation for the maintenance of standing genetic variation, and calls for a better understanding of the fitness consequences of adaptive variation across habitats and genomic backgrounds.

KEYWORDS
allele frequency, ancestor, evolutionary genomics, *Gasterosteus aculeatus*, migration–selection balance, North Uist, purifying selection, whole-genome sequencing

1 | INTRODUCTION

In eukaryotes, adaptation of populations to novel ecological conditions often occurs from standing genetic variation (SGV), that is, selectively relevant variation pre-existing in the ancestor (Barrett & Schluter, 2008; Hermisson & Pennings, 2005; Matuszewski et al., 2015; Messer & Petrov, 2013; Orr & Betancourt, 2001). A puzzle, however, is how SGV is maintained in the ancestor (Yeaman, 2015): if genetic variants are favoured by selection in a novel, derived habitat, should they not be unfavourable and hence eliminated by purifying selection in the ancestral habitat? One solution to this paradox is that genetic variants favoured in the derived habitat are maintained as SGV in the ancestor by continued hybridization (and hence gene flow) between derived and ancestral populations, thus counteracting the selective removal of these variants in the latter (Barrett & Schluter, 2008; Bolnick & Nosil, 2007; Colosimo et al., 2005; Galloway et al., 2020; Schluter & Conte, 2009; Yeaman & Whitlock, 2011). An alternative idea is that variants beneficial within the novel habitat are selectively neutral in the ancestral population when their frequency is relatively low. While this must obviously

hold for recessive variants (Barrett & Schluter, 2008), quantitative genetic models suggest that when the traits under selection are highly polygenic (i.e., influenced by a great number of loci), adaptive divergence may generally occur primarily via the establishment of linkage disequilibrium among alleles and involve only relatively subtle (or at least incomplete) allele frequency differentiation (Kremer & Le Corre, 2012; Latta, 1998; Le Corre & Kremer, 2012). In this case, SGV could persist in the ancestor simply because there is no purifying selection to complete its elimination. The relative importance of these two not mutually exclusive explanations for the maintenance of SGV, gene flow–selection balance and selective neutrality, remains unknown and has, to the best of our knowledge, not been subject to empirical investigation. An obstacle for doing so is that organismal systems are required in which adaptive genetic variation can be detected and quantified in both derived and ancestral populations simultaneously.

We here perform such an investigation in threespine stickleback fish (*Gasterosteus aculeatus*) by focusing on genetic variation promoting the adaptation of populations to acidic freshwater habitats after the recent (postglacial) colonization of these habitats by ancestral marine stickleback. Adaptation to acidic waters probably involves numerous traits, but particularly obvious elements include the reduction of external skeletal armour and body size in some acid-adapted stickleback populations relative to their ancestor (and to standard freshwater-adapted stickleback) (Figure 1a) (Bourgeois et al., 1994; Campbell, 1985; Giles, 1983; Haenel et al., 2019a; Klepaker et al., 2016; Magalhaes et al., 2016; Spence et al., 2013). The function of this evolution is likely to be reduced metabolic demands, conferring an advantage in nutrient-depleted acidic habitats. (Note that for simplicity, we will use the terms acidic habitats and acidic adaptation throughout this paper, but we acknowledge that selection may not necessarily be mediated by pH [alone], but by an associated shortage in dissolved ions.) Although marine threespine stickleback have colonized innumerable freshwater habitats across the northern hemisphere, morphological adaptation to acidic habitats is reported only from relatively few locations across the species' range (Campbell, 1985; Bourgeois et al., 1994; Klepaker et al.,2013). An exception is North Uist (Outer Hebrides, Scotland) (Figure 1b), an island on which acidic-adapted stickleback ecomorphs are common. Due to its particular surface geology (Waterston et al., 1979), the eastern part of this island harbours numerous acidic lakes (pH around 5–6) inhabited by archetypal acidic-adapted stickleback that have probably evolved multiple times independently (Giles, 1983; Haenel et al., 2019a; Klepaker et al., 2016; Magalhaes et al., 2016; Spence et al., 2013). This parallel evolution has occurred though the deterministic sorting of SGV available in the marine ancestor, because alleles recruited repeatedly for acidic adaptation are consistently found in extant marine stickleback breeding in coastal habitats of North Uist, albeit generally at modest to low frequency (Haenel et al., 2019a). What remains unknown is whether this SGV primarily reflects the continued flow of acid-favoured alleles into marine stickleback by hybridization, or whether alleles beneficial to acidic adaptation segregate largely neutrally at these frequencies in marine fish.

To address this question, we here use whole-genome sequence data to examine SGV in marine stickleback across the Atlantic Ocean. We hypothesize that if the presence of SGV relevant to acidic adaptation in marine stickleback around North Uist reflects a balance between gene flow and purifying selection, the frequency of alleles favoured in acidic habitats should be elevated in marine stickleback breeding around North Uist compared to marine stickleback sampled from more distant locations (Figure 2, top). The reason is that acidic lakes represent an uncommon freshwater habitat outside North Uist, and the acidic-adapted ecomorphs common on this island are rare on a worldwide basis. Purifying selection should therefore vastly outbalance the input of deleterious acidic-favoured alleles by hybridization in marine stickleback far from North Uist. Alternatively, the frequency of acidic-favoured alleles may not be elevated in marine stickleback breeding around North Uist compared to marine fish in general (Figure 2, bottom), suggesting that purifying selection against these alleles is weak or absent in marine stickleback at large. As we show, our data support this latter scenario, thus highlighting selective neutrality as an underappreciated explanation for the maintenance of SGV.

2 | MATERIAL AND METHODS

2.1 | Stickleback samples, DNA library preparation and sequencing

A precondition for our analysis of SGV in marine stickleback was the initial identification of genetic polymorphisms important to acidic adaptation. For this, we considered five acidic and five basic lakes from North Uist from which individual DNA was already available (Haenel et al., 2019a,2019b) (Figure 1b, Table S1). We refer to the latter habitat type as "basic" for terminological consistency with our previous work, but emphasize that the fish inhabiting these lakes represent the standard freshwater stickleback ecomorph widespread across the range of *Gasterosteus aculeatus*. We chose 20 individuals from each of these freshwater populations at random and combined their DNA to equal molarity without PCR (polymerase chain reaction)-enrichment into either an acidic or a basic pool of 100 individuals each. The goal of this pooling (and the subsequent pooled sequencing, hereafter poolSeq) was to obtain relatively precise allele frequency estimates for acidic versus basic stickleback in general, while ignoring allele frequencies within each specific population. To nevertheless have access to individual genotypes and haplotype information, we additionally chose two individuals from each acidic and basic population at random for individual sequencing (indSeq).

To explore the extent to which adaptive genetic variation discovered in freshwater fish is present as SGV in marine stickleback, we focused on samples from six locations across the Atlantic Ocean: North Uist (NU), Ireland (IR), The Netherlands (NL), Germany (DE), Iceland (IS) and Eastern Canada (CA) (Figure 1b; Table S1; note that North Uist subsumes two nearby marine sample sites, ARDH and

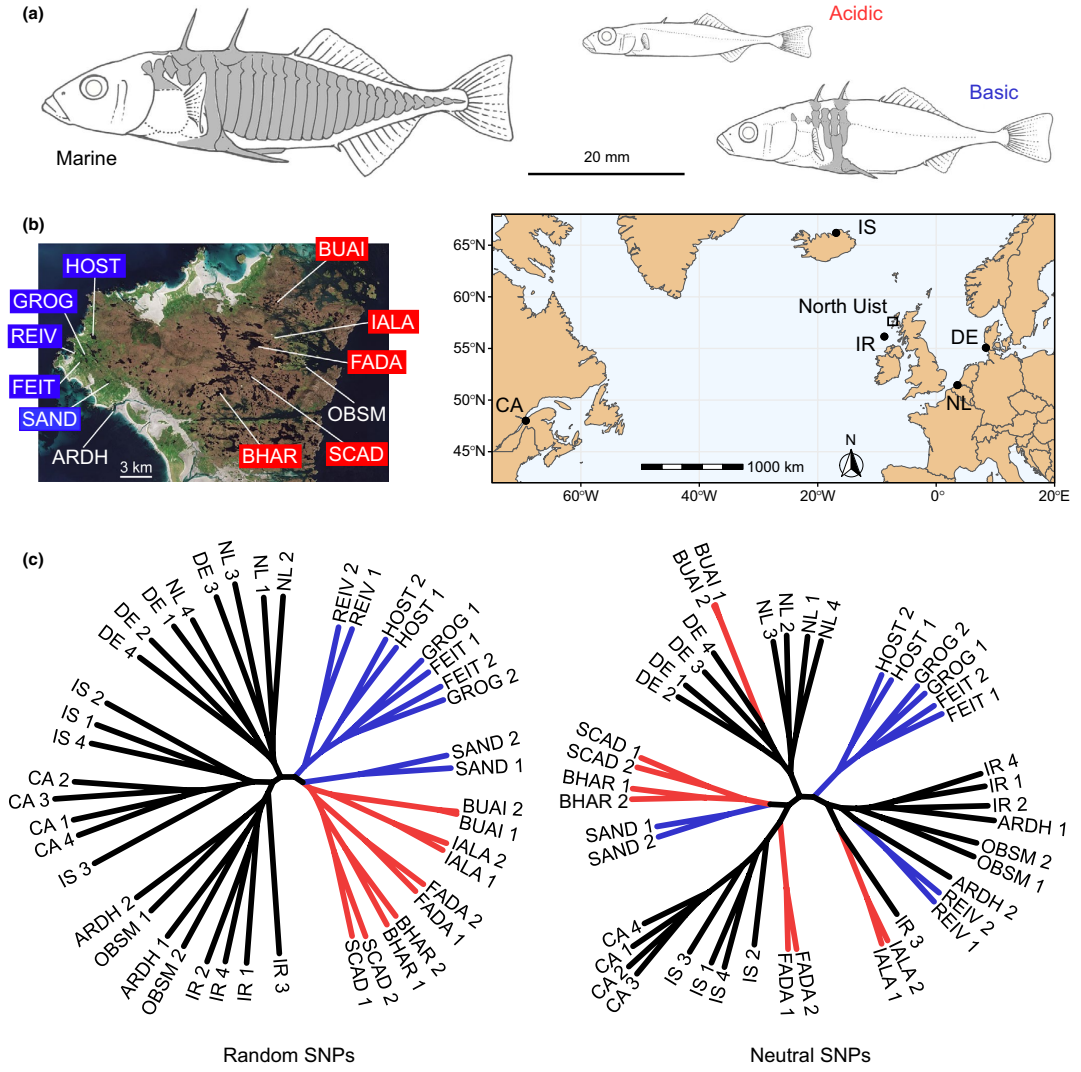


FIGURE 1 (a) Typical stickleback ecomorphs from marine, acidic freshwater and standard freshwater (here called "basic") habitats, highlighting the particularly strong reduction in bony armour and body size in acidic stickleback. Key external skeletal elements (dorsal spines, lateral plates, pelvic complex) are shaded in grey. (b) Image of North Uist (left), indicating the acidic (red) and basic (blue) lakes from which freshwater stickleback were sampled. The sites ARDH and OBSM represent locations at which marine stickleback were collected. The other five Atlantic marine sample sites are located in the map (right; North Uist is indicated by the small rectangle). (c) Unrooted maximum-likelihood phylograms showing the genetic similarity among 44 total marine, acidic and basic stickleback individuals. The left tree is based on 200,000 SNPs selected at random across the genome, whereas the right tree uses 120,448 SNPs filtered to be little influenced by selection (i.e., exhibiting low allele frequency differentiation in both marine–freshwater and acidic–basic genome scans, and located in chromosome regions showing high recombination rates)

OBSM). From each of these marine locations, we aimed for a sample size of around 25 individuals. Except for North Uist, from which marine individual-level whole-genome sequence data were already available (Haenel et al., 2019a,2019b), individual DNA was extracted using the Quick-DNA Miniprep Plus Kit (Zymo Research). For the estimation of population allele frequencies via poolSeq, individual DNA was then combined to equal molarity without PCR-enrichment within each of the five new locations. In addition, four individuals

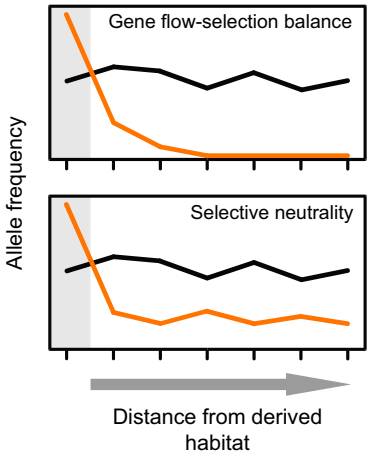


FIGURE 2 Two alternative explanations for the maintenance of adaptive standing genetic variation (SGV) in ancestral populations. Under gene flow–selection balance (top), genetic variants adaptive and hence at high frequency within a derived habitat (grey background shading) are unconditionally disfavoured in the ancestral habitat (white background shading). These variants, however, may still occur at appreciable frequency in the ancestral habitat if hybridization between populations from the two habitats leads to gene flow. A prediction based on this scenario is that if the opportunity for hybridization is geographically restricted, the frequency in the ancestral habitat of variants favoured in the derived habitat should decline with increasing distance from the derived habitat (orange curve; the ticks represent hypothetical sample sites) because purifying selection increasingly outbalances gene flow. Such spatial change in allele frequencies would not be expected at ecologically neutral polymorphisms (black curve). Under selective neutrality (bottom), we assume that alleles favoured in the derived habitat are selectively neutral within the ancestral habitat when their frequency is relatively low, thus allowing their persistence. The key prediction under this latter scenario is that the frequency in the ancestral habitat of variants favoured in the derived habitat does not decline with increasing geographical distance from the derived habitat

from each of these locations were chosen at random for indSeq (Table S1).

The 47 total DNA libraries (seven pools and 40 individuals) were paired-end sequenced to 150 bp together on a single S4 flow cell of an Illumina NovaSeq 6000 instrument, producing a genome-wide median read depth per base pair of 85× on average across the pools, and of 16× across the individuals (details given in Table S1).

2.2 | SNP discovery

Raw sequences reads (Haenel et al., 2019b,2021) were parsed by library (pool or individual) and aligned to the third-generation stickleback reference genome assembly (Glazer et al., 2015) by using NOVOALIGN (version 4.0, http://www.novocraft.com/products/novoalign/; alignment settings provided in the Supplementary Codes). From the alignments, we derived nucleotide counts (pileups) for all genome-wide positions by using the *pileup* function from the *Rsamtools* Rpackage (Morgan et al., 2017; unless specified otherwise, all analyses were implemented with the Rlanguage, version 3.6.0; RDevelopment Core Team, 2019). Single-nucleotide polymorphisms (SNPs) were then ascertained in two ways: for an initial exploration of population structure among our marine and freshwater samples, we used the pileup data derived from indSeq. Genomic positions qualified as SNPs if the minor allele frequency (MAF) was at least 0.04 across the 24 marine individuals (thus excluding positions appearing variable due to sequencing error only); if cumulative read depth across the marine fish was no greater than 1000 (thus effectively eliminating repeated genomic elements); if all 44 stickleback individuals displayed at least 1× read depth (thus excluding positions with missing data); and if the physical distance to the nearest SNP was at least 100 bp (thus ruling out SNP clusters caused by micro-indels). This stringent quality filtering resulted in our "indSeq SNPs" including 1.65 million markers across the 447-Mb stickleback genome. Analyses based on an alternative SNP panel (1.61 million SNPs) obtained by applying the MAF and cumulative read depth threshold to the 20 freshwater instead of the marine individuals consistently produced similar results (details not reported).

For the discovery of genetic variation important to acidic adaptation and the subsequent exploration of SGV, SNPs were ascertained based on the poolSeq data from the acidic and basic fish. We here required a read depth between 100 and 500× and a MAF of at least 0.25 across the two pools combined, and a read depth of at least 50× within each pool. The 1.5 million "poolSeq SNPs" passing these filters were genotyped in all freshwater and marine population pools separately.

2.3 | Population structure

As a first analytical step, we explored population structure based on genealogies derived from the indSeq SNPs. The purpose was to develop a sense for the genetic relatedness among marine stickleback across the Atlantic Ocean, and to reassess the relatedness of the freshwater populations among each other and to marine fish based on SNP data from whole-genome indSeq (in Haenel et al., 2019a the latter was done with SNPs derived from pooled RADseq [restriction site-associated DNA sequencing]). For computational efficiency, we reduced the full indSeq SNP panel to a random subset of 200,000 autosomal SNPs, additionally considering sample sizes of 100,000 and 15,000 SNPs in supplementary analyses (all these data sets were largely independent, as the choice of SNPs was random). For all 44 marine and freshwater individuals, we then derived haploid multilocus genotypes by drawing at each SNP the more frequent allele, or a random allele when both were equally frequent. This haploid strategy (Berner, 2021) circumvented the

ambiguity of diploid genotyping in individuals with low read depth. The haploid genotypes were then concatenated to nucleotide strings in fasta format.

The genotype data above were derived from SNPs chosen at random across the genome. However, both marine–freshwater and acidic–basic divergence in stickleback involves selection on numerous loci across the genome (Bassham et al., 2018; Fang et al., 2020; Haenel et al., 2019a; Jones, Grabherr, et al., 2012; Roesti et al., 2014; Terekhanova et al., 2019). To assess to what extent natural selection influences population structure, we additionally explored the genetic relatedness among our marine and freshwater individuals based on a subset of indSeq SNPs filtered to reduce the influence of selection. Following the strategy of Haenel et al. (2019a), we excluded SNPs exhibiting an absolute allele frequency difference (*AFD*; Berner, 2019) >0.4 in both a global marine–freshwater comparison performed by pooling two random nucleotides drawn from the pileup of each individual at each SNP within the marine vs. freshwater group of individuals, and in the acidic–basic comparison described below. As the latter included an MAF threshold of 0.25, we applied the same threshold in the marine–freshwater comparison. Moreover, we here considered exclusively SNPs located within the peripheral 5 Mb of each chromosome (Berner & Roesti, 2017). These regions display particularly high recombination rates in stickleback (Glazer et al., 2015; Roesti et al., 2013), and hence are those least affected by hitchhiking (linked selection). The 120,448 SNPs passing these filters were treated as above to obtain haploid genotype strings. We hereafter call the randomly chosen genotype data "Random SNPs" and the markers chosen to reduce the footprint of selection "Neutral SNPs", emphasizing that in the latter, a signal of selection may still persist.

For an earlier investigation of the genetic relatedness among North Uist stickleback based on poolSeq data, we used synthetic multilocus genotypes generated by concatenating alleles drawn from RAD-sequenced sample pools (Haenel et al., 2019a), thereby erasing individual-level haplotype structure. To assess the value of such synthetic genotypes for capturing genetic structure among populations, we here pooled the nucleotide counts at a number of random and neutral SNPs matching the individual-level data described above. We then drew a single nucleotide per sample location according to the observed pooled allele frequencies, and saved these draws concatenated to a single haploid nucleotide string per location in fasta format. The synthetic genotype data produced in this way allowed comparing genealogies based on truly individual-aware vs. synthetic genotypes derived from the same SNP panel.

Based on the genotype files, genealogies were generated by using the *ape* (version 5; Paradis & Schliep, 2018) and *phangorn* (version 2.5.5; Schliep, 2011) R packages. We determined the most appropriate models of sequence evolution (mostly GTR+G), constructed maximum-likelihood genealogies, and visualized them as unrooted phylograms. Node support was determined based on 500 bootstrap iterations. As an alternative to phylograms, we also considered exploring population structure by ordination (principal coordinates analysis). However, the proportion of variation captured by the first ordination axes was consistently small (~8% or less). We therefore considered ordination an ineffective tool for pattern recognition.

2.4 | Identifying alleles important to acidic adaptation, and quantifying their frequencies in marine stickleback

To identify alleles important to the adaptation of stickleback to acidic habitats, we performed genome-wide differentiation mapping between the acidic and basic sample pools. That is, we scanned the poolSeq SNPs for positions exhibiting extremely high global differentiation between stickleback from acidic vs. basic lakes. The reason why we did not define genetic variation important for acidic adaptation simply as SNPs highly differentiated between acidic and marine fish is that this would mostly have uncovered genetic variation important to marine–freshwater divergence in general. Such variation is abundant in North Uist stickleback (Figure S3 in Haenel et al., 2019a; see also Jones, Grabherr, et al., 2012; Roesti et al., 2014; Bassham et al., 2018; Fang et al., 2020; Terekhanova et al., 2019). Our focus, however, was specifically on genetic variation for which gene flow into marine fish must be rare and geographically restricted. Acidic–basic differentiation was expressed by the absolute allele frequency difference *AFD*. Positions qualified as high-differentiation SNPs if they showed $AFD \geq 0.85$, approximately corresponding to the top 0.01 percent of the *AFD* distribution. This *AFD* threshold was more stringent than in Haenel et al. (2019a) (0.70) because a higher marker resolution was available, and was chosen to maximize the strength of acidic–basic differentiation while still yielding an adequate number of SNPs for downstream analyses. The positions were further required to be autosomal, and to be physically separated from each other by at least 100 kb to ensure independence (tight linkage disequilibrium typically decays over much shorter distances in stickleback; e.g., Roesti et al., 2015). With these criteria, we obtained a panel of 50 "adaptive SNPs", that is, positions at which one allele appears strongly and consistently selectively favoured in acidic habitats. As a basis for comparison, we analogously selected a panel of 500 "baseline SNPs" from the same genome scan. These latter polymorphisms were also required to be separated by at least 100 kb, but to exhibit minimal differentiation (*AFD* within 0.1% of the genome-wide median) between the acidic and the basic pool. The latter criterion ensured that these SNPs did not tag genome regions (consistently) involved in acidic adaptation. At each of the adaptive SNPs, we then defined the nucleotide predominant in the acidic pool as the "acidic allele," and determined and graphed the frequency of these alleles in all six marine sample pools. An analogous analysis was performed for the baseline SNPs, here defining the acidic allele as the one relatively more common in the acidic than the basic pool. Our prediction was that if genetic variation at the adaptive SNPs in marine stickleback reflects gene flow–selection balance, the frequency of the acidic alleles at these markers (but not at the baseline SNPs) should be elevated in marine stickleback sampled on North

Uist. As a resource, we additionally compiled all genes located within a 100-kb window centred at each adaptive SNP.

For three exemplary adaptive SNPs, we further visualized the diversity and distribution of surrounding haplotypes among our samples based on haplotype networks. The markers chosen included the adaptive SNP exhibiting the strongest acidic–basic differentiation in the present study (*AFD* = 0.96), the adaptive SNP tagging the genome region showing the strongest acidic–basic differentiation in a previous investigation (Figure 3a in Haenel et al., 2019a), and the adaptive SNP located on a known inversion polymorphism (Haenel et al., 2019a; Jones, Grabherr, et al., 2012; Roesti et al., 2015). Using the raw nucleotide counts derived from indSeq, we performed individual diploid genotyping for all nucleotide positions exhibiting a read depth of 10× or greater across a 5-kb window centred on the adaptive SNPs, considering positions as heterozygous if their MAF was >0.1. Individuals with >25% missing genotypes were omitted. Based on the remaining data, positions qualified as informative SNPs if they displayed ≤40% missing genotypes and a MAF of at least 0.05. The resulting genotype matrices were subjected to phasing with FASTPHASE version 1.4.8 (Scheet & Stephens, 2006; settings provided in the Supplementary Codes). Haplotype genealogies were then constructed with RAXML version 8 (Stamatakis, 2014) and visualized as haplotype networks in FITCHI (Matschiner, 2016) (settings provided in the Supplementary Codes).

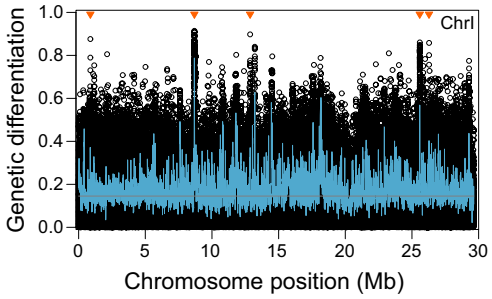


FIGURE 3 Genetic differentiation, quantified by the absolute allele frequency difference *AFD*, between the acidic and basic stickleback pool along an exemplary chromosome. The black circles represent individual SNPs, the blue curve shows average differentiation across sliding windows of 10 kb with 5-kb overlap (windows with fewer than six SNPs were discarded), and the grey line gives the genome-wide median differentiation (0.145). The orange triangles denote the adaptive SNPs on this chromosome; that is, the markers exhibiting extremely strong and consistent acidic–basic differentiation used to explore adaptive standing genetic variation in marine stickleback

3 | RESULTS AND DISCUSSION

3.1 | Population structure

Our high-resolution SNP genealogies revealed consistent yet modest genetic structure among marine stickleback from the Atlantic. Specifically, the phylograms based on SNPs both chosen randomly across the genome and filtered stringently to reduce the influence of selection recovered three marine branches (Figure 1c; bootstrap support is given in Figure S1). These branches were formed by the marine individuals from North Uist and Ireland (ARDH, OBSM, IR), the two samples from the North Sea (DE, NL), and stickleback from Canada and Iceland (CA, IS). Within these branches, however, marine fish from a given location generally did not emerge as monophyletic, except for the Canadian individuals collected thousands of kilometres from the nearest sampling locations (IR, IS) (Figure 1b). In contrast to the marine fish, our freshwater samples exhibited genetic structure differing fundamentally between the random and neutral SNP panels (Figure 1c). Based on the former, all freshwater stickleback together grouped to a single, well-supported branch distinct from marine fish, and within this freshwater branch, individuals clustered almost perfectly according to acidic vs. basic habitat. This ecological structure largely vanished when using SNPs ascertained to reduce the influence of selection. Moreover, contrary to marine stickleback, freshwater individuals almost consistently grouped by sampling location, despite the dramatically smaller geographical distance among the lakes compared to the marine locations (Figure 1b). All these patterns remained qualitatively consistent when using sparser data sets, and when replacing individual-level by synthetic genotypes derived from pooled data (Figure S1). The latter confirms that poolSeq data enable meaningful genealogical analyses at the population level (Haenel et al., 2019a).

The modest genetic structure among our marine locations within the three marine branches is consistent with the notion that marine stickleback display large population sizes, and that genetic drift is relatively weak (Catchen et al., 2013; Hohenlohe et al., 2010; Jones, Chan, et al., 2012; Lescak et al., 2015; Mäkinen et al., 2006; Roesti et al., 2014). This view is also well supported by the comparison of genetic differentiation among marine vs. among freshwater samples: while genome-wide median *AFD* was 0.132 across all pairwise marine sample comparisons (0.019 when expressed by F_{ST}, Nei, 1973; individual values are presented in Table S2), much higher values were observed across the pairwise comparisons between freshwater populations (*AFD* = 0.219, F_{ST} =0.068). (The latter values were derived from differentiation data presented in Table S2 of Haenel et al., 2019a; indSeq performed for the present study included too few individuals per population, and poolSeq used combinations of individuals from multiple populations, both precluding the reliable estimation of population differentiation.) Given weak drift in marine stickleback, we expect that deleterious genetic variation introduced by hybridization with freshwater fish should be eliminated efficiently. Nevertheless, stickleback across the Atlantic clearly do exhibit genetic structure related to geography. Assuming gene

flow–selection balance as a cause for the maintenance of SGV, we would therefore expect differences in the level of SGV among broad regions within the Atlantic if these regions differed in the input of maladaptive acidic alleles. A further insight into marine stickleback emerging from both the random and neutral SNPs is that the freshwater populations from North Uist are genetically no more similar to marine fish sampled in immediate (ARDH, OBSM) or relative (IR) proximity than to the samples from the much more distant marine locations. This implies that at the genome-wide level, any Atlantic marine sample—irrespective of its precise geographical origin (and including offshore samples such as IR; Table S1)—serves as an adequate representation of ancestral Atlantic marine stickleback (see also Kirch et al., 2021).

An intriguing finding emerging from the genealogy is the nearly perfect segregation of stickleback by habitat when using SNPs sampled at random across the genome. At first glance, this may stimulate the interpretation that on North Uist, initially a single freshwater stickleback form evolved, subsequently differentiated into a single acidic and basic ecomorph, and these ecomorphs then split into multiple subpopulations. Apart from being hydrogeographically implausible (see the Supporting Discussion in Haenel et al., 2019a), this interpretation is challenged by the genetic structure revealed by the neutral SNPs: the deep separation of freshwater populations on North Uist based on this marker panel indicates that acidic and basic ecomorphs have arisen multiple times independently through the adaptive sorting of ancestral marine SGV (Magalhaes et al., 2016; Haenel et al., 2019a; see also Bell et al., 1993). The contrasting results obtained from random vs. neutral SNPs in freshwater but not marine stickleback highlight, on the one hand, how deterministically genome-wide polygenic selection and associated hitchhiking during freshwater adaptation can shape genetic population structure and thus confound neutral evolutionary history (see also Berner, 2021; Berner & Roesti, 2017). On the other hand, these results indicate that the genomes of stickleback populations recently adapted to ecologically novel freshwater habitats are much more profoundly shaped by selection than the genomes of the ancestral marine form. Nevertheless, the deep separation among the freshwater populations observed in both types of genealogies (and mirrored by genome-wide differentiation; Table S2 in Haenel et al., 2019a) make clear that drift associated with relatively small population size has also played a fundamental role in the evolution of our acidic and basic stickleback populations.

3.2 | Loci important to acidic adaptation and their allele frequencies across Atlantic stickleback

Our analysis of genetic structure revealed striking genome-wide evidence of selection, including between acidic and basic ecomorphs. To investigate how polymorphisms important to acidic adaptation are maintained as SGV in marine stickleback, we searched for loci consistently involved in acidic adaptation based on the genome-wide comparison of acidic vs. basic poolSeq data (Figure 3; differentiation profiles across all chromosomes are presented in Figure S2). This identified 50 independent adaptive SNPs nearly fixed for alternative alleles between the two freshwater ecomorphs (*AFD* 0.851–0.960; genome-wide median differentiation was 0.145) (Figure 4a; all adaptive SNPs are characterized in Table S3, and associated genes listed in Table S4). These adaptive SNPs recovered many of the genome regions identified as important to acidic–basic differentiation in Haenel et al. (2019a), based on partly independent specimen panels and a different analytical approach. Specifically, 15 of the 19 regions of highest acidic–basic differentiation inferred in Haenel et al., 2019a (i.e., the regions containing the "top core SNPs" in that study) also exhibited a marker qualifying as adaptive SNP in the present investigation (Figure 4a; Figure S3). However, given the much higher (whole-genome) marker resolution, the present study also identified numerous novel regions (Figure 4a; Figure S2). Haplotype networks derived from genotypes phased across 5 kb around three exemplary adaptive SNPs indicated that these markers generally represent longer DNA tracts differentiated between the ecomorphs (Figure 4b). Across these exemplary regions, acidic stickleback populations generally shared closely related haplotypes distinct from the haplotypes prevailing in marine (and basic) fish, although sometimes acidic individuals exhibited marine haplotypes (chromosome IX and XI) and vice versa (chromosome XI).

At the adaptive SNPs, marine stickleback generally exhibited lower frequencies for the alleles characteristic of acidic fish (acidic alleles; median frequency across all SNP by marine sample combinations: 0.30) than for the alleles typical of the basic populations (median frequency 0.70) (Figure 5a; Table S3). Also, the acidic alleles occurred at a lower overall frequency at the adaptive SNPs than at the baseline SNPs not under consistent acidic–basic divergence (median frequency across all baseline SNPs by marine sample combinations: 0.46). A few adaptive SNPs, however, were exceptional in that the acidic allele occurred at consistently high frequency, or even close to fixation, in the ocean (e.g., the SNPs 8, 10 and 28 in Table S3; an exemplary haplotype network for such a SNP is shown in Figure S4). These polymorphisms thus made it into our panel of adaptive SNPs because of massive allele frequency shifts during the adaptation to the basic but not to the acidic habitats.

Overall, these findings are in line with observations in Haenel et al. (2019a) and indicate that alleles presumably important for the adaptation to ecologically highly derived acidic habitats tend to be unfavourable in ancestral marine stickleback when occurring at high frequency. Interestingly, however, we found no indication that the frequency of the acidic allele at the adaptive SNPs was elevated in marine samples collected around North Uist compared to samples from more distant locations (Figure 5a; compatibility intervals for the median frequency of the acidic alleles for all samples are presented in Figure S5); the frequency of these alleles was highly stable across all our marine samples. This key finding was reproduced when considering exclusively the subset of adaptive SNPs at which the acidic allele proved the minor allele within *all* marine samples ($n = 21$; indicated in Table S3) (Figure 5b; Figure S5; median frequency across all SNPs by marine sample combinations:

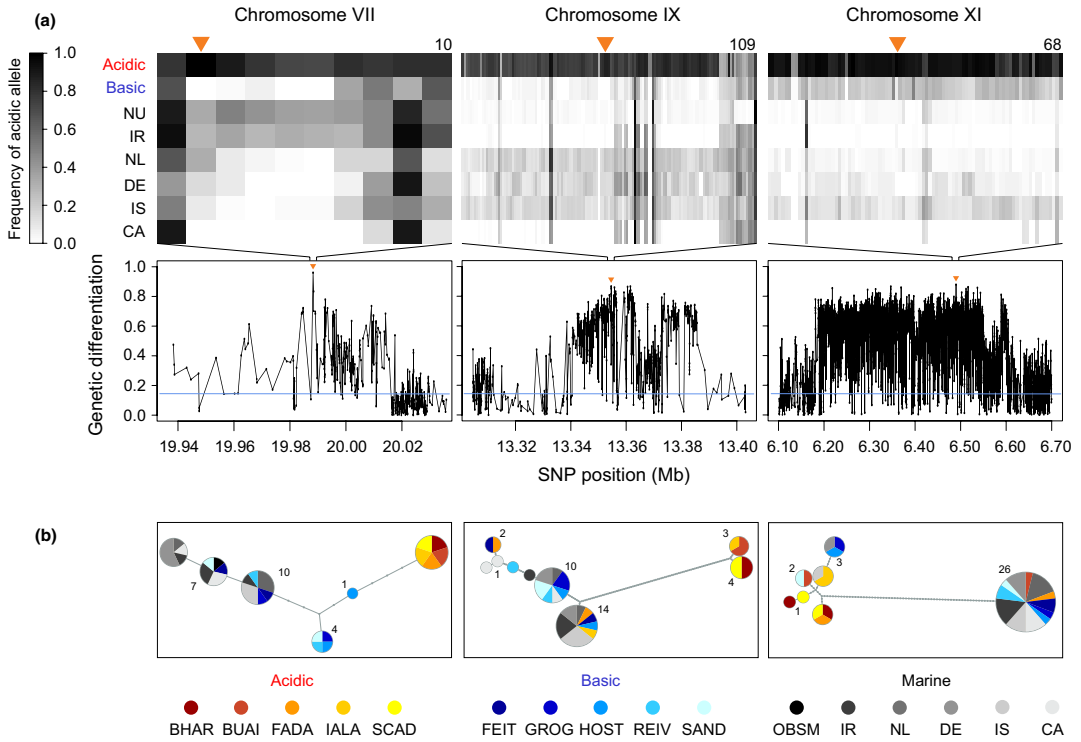


FIGURE 4 Loci important to acidic adaptation, and their allele frequencies and haplotypes across samples. The lower panels in (a) show three exemplary genome regions exhibiting strong differentiation between the acidic and basic stickleback pools. The dots connected by lines represent individual SNPs, and the horizontal blue line indicates genome-wide median differentiation. The markers exhibiting the highest differentiation in these regions are marked by orange triangles and were included in the panel of adaptive SNPs (*AFD* ≥ 0.85). The adaptive SNP on chromosome VII is the most strongly differentiated marker in our study, while the locus on chromosome IX showed the strongest acidic–basic differentiation in a previous genome scan (Figure 3A in Haenel et al., 2019a). The locus on chromosome XI is an inversion. The width of the visualized chromosome window is 100 kb for the loci on chromosomes VII and IX, and 600 kb for the inversion locus. The upper panels in (a) indicate for each freshwater and marine stickleback pool the frequency of the allele predominant in the acidic pool (acidic allele) at all SNPs within a 5-kb window centred at the three adaptive SNPs. Each SNP is a separate column, and the number of SNPs is indicated on the top right of each panel. The NU pool combines marine individuals from the North Uist sites ARDH and OBSM. (b) Haplotype genealogies based on phased genotypes derived from individual sequencing at SNPs across the same 5-kb windows. Pies represent unique haplotypes and edges connecting pies or nodes indicate one inferred mutational step. Within each panel, sample size is given for one pie per size class. Note that the acidic populations generally share haplotypes highly distinct from those prevailing in the marine samples and in the basic populations

0.10); that is, the subset of markers at which purifying selection in marine stickleback appears particularly plausible because acidic adaptation involves a particularly strong shift away from the ancestral allele frequency.

The finding of similar frequencies of alleles important to adaptation to acidic waters across Atlantic marine stickleback challenges perpetual antagonism between gene flow and purifying selection (Bassham et al., 2018; Galloway et al., 2020; Schluter & Conte, 2009) as a sufficient explanation for the maintenance of adaptive SGV in the ocean. Instead, we propose that acidic alleles can persist neutrally in marine populations when occurring at moderate to low frequencies. Purifying selection certainly plays a role, but primarily by impeding these alleles from rising to high frequency in marine stickleback. Note that the average frequency of the acidic alleles in the ocean was still around 0.3 (Figure 5a; Figure S5); at many adaptive loci, a substantial proportion of marine stickleback are thus expected to be homozygous for the acidic allele, so that purifying selection should still be effective even when these alleles were recessive. We therefore argue that the reason for the persistence of acidic alleles in marine populations is not their recessivity, but their selective neutrality when relatively uncommon. This interpretation supports quantitative genetic models under which polygenic adaptation can be achieved by moderate allele frequency shifts (Kremer & Le Corre, 2012; Latta, 1998; Le Corre & Kremer, 2012).

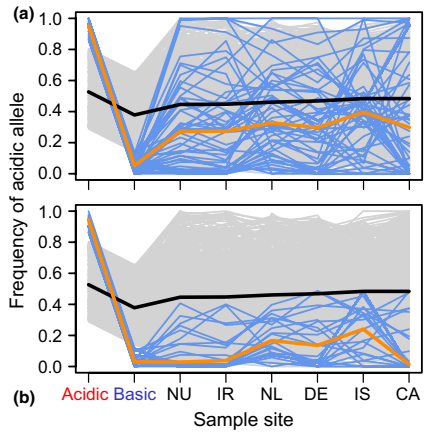


FIGURE 5 Frequency of the acidic allele at the adaptive and baseline SNPs. (a) The blue lines give the frequency of the acidic allele at each of the 50 adaptive SNPs in each sample pool, and the orange line indicates the median frequency. The grey lines show the acidic allele frequency at 500 baseline SNPs exhibiting a magnitude of acidic–basic differentiation near the genome-wide median (their median frequency is indicated by the black line). The first two sites from the left are the freshwater pools from North Uist used to identify the adaptive SNPs. The other locations represent marine stickleback (NU combines individuals from the marine North Uist samples ARDH and OBSM). The marine locations are ordered by increasing approximate swimming distance from North Uist. Note that the subtle allele frequency differentiation between the acidic and basic pool at the baseline SNPs is expected technically because at these markers too, the acidic allele was defined as the one relatively more frequent in the acidic than the basic pool. Panel (b) follows the same format as (a) but shows data only for the subset of adaptive SNPs at which the acidic allele is the minor allele within all marine sample pools. Both graphs convey that the frequency of alleles important to acidic adaptation is not elevated in marine stickleback close to North Uist than further away

An important caveat to consider is that although acidic habitats and the associated stickleback ecomorphs (Figure 1a) are exceptionally common on North Uist and rare elsewhere (Campbell, 1985; Bourgeois et al., 1994; Klepaker et al. 2013), the potential of marine stickleback to hybridize with acidic-adapted freshwater populations was not explicitly manipulated or controlled among our Atlantic marine samples. Is it plausible that gene flow from acidic-adapted to marine stickleback is more widespread than we assume, sufficiently so to raise acidic alleles to substantial frequencies in marine stickleback all across the Atlantic despite purifying selection? In our view, the marine samples from the North Sea (DE, NL) refute this concern: western mainland Europe is densely populated and its Ichthyofauna is well investigated, but acidic stickleback ecomorphs have to our knowledge not been reported. Gene flow of acidic alleles into marine fish thus appears highly unlikely across this region, and yet the frequencies of acidic alleles are not reduced in these specific marine samples (Figure 5; Figure S5), consistent with the selective neutrality of these alleles when occurring at the frequencies observed in marine fish. Similar reasoning applies to marine stickleback around Iceland, because highly acidic freshwater habitats seem to be absent in Iceland (Magalhaes et al., 2021).

4 | CONCLUSIONS

Adaptation commonly occurs from standing genetic variation, but how this variation is maintained in ancestral populations is little explored. We have here presented observational evidence suggesting that, overall, genetic variants important to adaptation to a highly derived habitat are maintained at moderate frequencies within the ancestral habitat. These variants do not appear to occur in higher frequencies in geographical regions where ancestral populations have a higher opportunity for gene flow from derived populations. We thus conclude that long-term gene flow–selection balance is an incomplete explanation for the maintenance of SGV. Instead, we propose that purifying selection of these variants in the ancestral habitat subsides as their frequency decreases, thus allowing their neutral persistence. This novel perspective on the maintenance of SGV should now be scrutinized by controlled experimental work quantifying the fitness consequences of individual genetic variants across different habitats and genomic backgrounds.

ACKNOWLEDGEMENTS

We thank the Swiss National Science Foundation (SNF grant 31003A_165826) for financial support to D.B.; the Freiwillige Akademische Gesellschaft Basel (FAG) for financial support to Q.H.; Louis Bernatchez, Jenny Boughman, Jacquelin DeFaveri, Bart Hellemans, Jun Kitano, Joost Raeymaekers, Mark Ravinet and Florent Sylvestre for providing marine stickleback samples; Walter Salzburger for sharing wet laboratory infrastructure; Brigitte Aeschbach and Nicolas Boileau for facilitating laboratory work; Christian Beisel, Ina Nissen-Naidanow and Elodie Vogel Burcklen for Illumina sequencing at the Quantitative Genomics Facility, D-BSSE, ETH Zürich; the developers of Novocraft for sharing their sequence aligner; Nicolás Lichilín Ortiz for help with cluster scripting; and three reviewers and Katie Lotterhos for constructive feedback on the manuscript. Computation was performed at the sciCORE scientific computing centre of the University of Basel (https://scicore.unibas.ch).

CONFLICT OF INTEREST

The authors declare no competing interests.

AUTHOR CONTRIBUTIONS

D.B. and Q.H. conceived the study; A.M. provided all freshwater and marine samples from North Uist; Q.H. performed wet laboratory work; Q.H., D.B. and L.G. wrote code and analysed genomic data; Q.H. and D.B. interpreted the results and wrote the manuscript.

DATA AVAILABILITY STATEMENT

Raw Illumina sequences for all individuals and pools are available from the NCBI Sequence Read Archive under BioProject no. PRJNA485717 (indSeq data from ARDH and OBSM), and from the European Nucleotide Archive under project no. PRJEB42736 (all other data). All code used for data analysis is provided as Supplementary Code in the Supporting Information.

ORCID

Andrew D. C. MacColl https://orcid.org/0000-0003-2102-6130
Daniel Berner https://orcid.org/0000-0003-3480-9046

REFERENCES

Barrett, R. D. H., & Schluter, D. (2008). Adaptation from standing genetic variation. *Trends in Ecology & Evolution*, *23*, 38–44.

Bassham, S., Catchen, J., Lescak, E., von Hippel, F. A., & Cresko, W. A. (2018). Repeated selection of alternatively adapted haplotypes creates sweeping genomic remodeling in stickleback. *Genetics*, *209*(3), 921–939. https://doi.org/10.1534/genetics.117.300610

Bell, M. A., Orti, G., Walker, J. A., & Koenings, J. P. (1993). Evolution of pelvic reduction in threespine stickleback fish: a test of competing hypotheses. *Evolution*, *47*, 906–914. https://doi.org/10.2307/2410193

Berner, D. (2019). Allele frequency difference *AFD* - an intuitive alternative to FST for quantifying genetic population differentiation. *Genes*, *10*(4), 308.

Berner, D. (2021). Re-evaluating the evidence for facilitation of stickleback speciation by admixture in the Lake Constance basin. *Nature Communications*, *12*, 2806.

Berner, D., & Roesti, M. (2017). Genomics of adaptive divergence with chromosome-scale heterogeneity in crossover rate. *Molecular Ecology*, *26*, 6351–6369.

Bolnick, D. I., & Nosil, P. (2007). Natural selection in populations subject to a migration load. *Evolution*, *61*(9), 2229–2243.

Bourgeois, J. F., Blouw, D. M., Koenings, J. P., & Bell, M. A. (1994). Multivariate analysis of geographic covariance between phenotypes and environments in the threespine stickleback, *Gasterosteus aculeatus*, from the Cook Inlet area, Alaska. *Canadian Journal of Zoology*, *72*, 1497–1509. https://doi.org/10.1139/z94-198

Campbell, R. N. (1985). Morphological variation in the three-spined stickleback (*Gasterosteus aculeatus*) in Scotland. *Behaviour*, *93*, 161–168. https://doi.org/10.1163/156853986X00838

Catchen, J., Bassham, S., Wilson, T., Currey, M., O'Brien, C., Yeates, Q., & Cresko, W. A. (2013). The population structure and recent colonization history of Oregon threespine stickleback determined using restriction-site associated DNA-sequencing. *Molecular Ecology*, *22*(11), 2864–2883. https://doi.org/10.1111/mec.12330

Colosimo, P. F., Hosemann, K. E., Balabhadra, S., Villarreal, G., Dickson, M., Grimwood, J., Schmutz, J., Myers, R. M., Schluter, D., & Kingsley, D. M. (2005). Widespread parallel evolution in sticklebacks by repeated fixation of ectodysplasin alleles. *Science*, *307*, 1928–1933.

Fang, B., Kemppainen, P., Momigliano, P., Feng, X., & Merilä, J. (2020). On the causes of geographically heterogeneous parallel evolution in sticklebacks. *Nature Ecology & Evolution*, *4*(8), 1105–1115. https://doi.org/10.1038/s41559-020-1222-6

Galloway, J., Cresko, W. A., & Ralph, P. (2020). A few stickleback suffice for the transport of alleles to new lakes. *G3: Genes, Genomes Genetics*, *10*(2), 505–514. https://doi.org/10.1101/713040

Giles, N. (1983). The possible role of environmental calcium levels during the evolution of phenotypic diversity in Outer Hebridean populations of the three-spined stickleback, *Gasterosteus aculeatus*. *Journal of Zoology*, *199*, 535–544.

Glazer, A. M., Killingbeck, E. E., Mitros, T., Rokhsar, D. S., & Miller, C. T. (2015). Genome assembly improvement and mapping convergently evolved skeletal traits in sticklebacks with genotyping-by-sequencing. *G3-Genes Genomes. Genetics*, *5*(7), 1463–1472. https://doi.org/10.1534/g3.115.017905

Haenel, Q., Guerard, L., MacColl, A. D. C., & Berner, D. (2021). *Individual whole-genome sequence data for North Uist freshwater and Atlantic marine stickleback*. https://www.ebi.ac.uk/ena/browser/view/PRJEB42736

Haenel, Q., Roesti, M., Moser, D., MacColl, A. D. C., & Berner, D. (2019a). Predictable genome-wide sorting of standing genetic variation during parallel adaptation to basic versus acidic environments in stickleback fish. *Evolution Letters*, *3*, 28–42.

Haenel, Q., Roesti, M., Moser, D., MacColl, A. D. C., & Berner, D. (2019b). *Individual whole-genome sequence data for North Uist marine stickleback*. https://www.ncbi.nlm.nih.gov/sra/?term=PRJNA485717

Hermisson, J., & Pennings, P. S. (2005). Soft sweeps: Molecular population genetics of adaptation from standing genetic variation. *Genetics*, *169*(4), 2335–2352.

Hohenlohe, P. A., Bassham, S., Etter, P. D., Stiffler, N., Johnson, E. A., & Cresko, W. A. (2010). Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLoS Genetics*, *6*(2), e1000862. https://doi.org/10.1371/journal.pgen.1000862

Jones, F., Chan, Y., Schmutz, J., Grimwood, J., Brady, S., Southwick, A., Absher, D., Myers, R., Reimchen, T., Deagle, B., Schluter, D., & Kingsley, D. (2012). A genome-wide SNP genotyping array reveals patterns of global and repeated species-pair divergence in sticklebacks. *Current Biology*, *22*(1), 83–90. http://linkinghub.elsevier.com/retrieve/pii/S0960982211013273

Jones, F. C., Grabherr, M. G., Chan, Y. F., Russell, P., Mauceli, E., Johnson, J., Swofford, R., Pirun, M., Zody, M. C., White, S., Birney, E., Searle, S., Schmutz, J., Grimwood, J., Dickson, M. C., Myers, R. M., Miller, C. T., Summers, B. R., Knecht, A. K., … Kingsley, D. M. (2012). The genomic basis of adaptive evolution in threespine sticklebacks. *Nature*, *484*(7392), 55–61. https://doi.org/10.1038/nature10944

Kirch, M., Romundset, A., Gilbert, M. T. P., Jones, F. C., & Foote, A. D. (2021). Ancient and modern stickleback genomes reveal the demographic constraints on adaptation. *Current Biology*, *31*(9), 2027–2036. https://doi.org/10.1016/j.cub.2021.02.027

Klepaker, T., Østbye, K., & Bell, M. A. (2013). Regressive evolution of the pelvic complex in stickleback fishes: A study of convergent evolution. *Evolutionary Ecology Research.*, *15*, 413–435.

Klepaker, T., Ostbye, K., Spence, R., Warren, M., Przybylski, M., & Smith, C. (2016). Selective agents in the adaptive radiation of Hebridean sticklebacks. *Evolutionary Ecology Research*, *17*, 243–262.

Kremer, A., & Le Corre, V. (2012). Decoupling of differentiation between traits and their underlying genes in response to divergent selection. *Heredity*, *108*(4), 375–385. https://doi.org/10.1038/hdy.2011.81

Latta, R. G. (1998). Differentiation of allelic frequencies at quantitative trait loci affecting locally adaptive traits. *American Naturalist*, *151*(3), 283–292. https://doi.org/10.1086/286119

Le corre, V., & Kremer, A. (2012). The genetic differentiation at quantitative trait loci under local adaptation. *Molecular Ecology*, *21*(7), 1548–1566. https://doi.org/10.1111/j.1365-294X.2012.05479.x

Lescak, E. A., Bassham, S. L., Catchen, J., Gelmond, O., Sherbick, M. L., von Hippel, F. A., & Cresko, W. A. (2015). Evolution of stickleback in 50 years on earthquake-uplifted islands. *Proceedings of the National Academy of Sciences of the United States of America*, *112*, E7204–E7212. https://doi.org/10.1073/pnas.1512020112

Magalhaes, I. S., D'Agostino, D., Hohenlohe, P. A., & MacColl, A. D. C. (2016). The ecology of an adaptive radiation of three-spined stickleback from North Uist. *Scotland. Mol. Ecol*, *25*(17), 4319–4336. https://doi.org/10.1111/mec.13746

Magalhaes, I. S., Whiting, J. R., D'Agostino, D., Hohenlohe, P. A., Mahmud, M., Bell, M. A., Skúlason, S., & MacColl, A. D. C. (2021). Intercontinental genomic parallelism in multiple three-spined

stickleback adaptive radiations. *Nat. Ecol. Evol*, *5*(2), 251–261. https://doi.org/10.1038/s41559-020-01341-8

Mäkinen, H. S., Cano, J. M., & Merilä, J. (2006). Genetic relationships among marine and freshwater populations of the European three-spined stickleback (*Gasterosteus aculeatus*) revealed by microsatellites. *Molecular Ecology*, *15*(6), 1519–1534. https://doi.org/10.1111/j.1365-294X.2006.02871.x

Matschiner, M. (2016). Fitchi: Haplotype genealogy graphs based on the Fitch algorithm. *Bioinformatics*, *32*(8), 1250–1252. https://doi.org/10.1093/bioinformatics/btv717

Matuszewski, S., Hermisson, J., & Kopp, M. (2015). Catch me if you can: Adaptation from standing genetic variation to a moving phenotypic optimum. *Genetics*, *200*(4), 1255–1274. https://doi.org/10.1534/genetics.115.178574

Messer, P. W., & Petrov, D. A. (2013). Population genomics of rapid adaptation by soft selective sweeps. *Trends in Ecology & Evolution*, *28*(11), 659–669. https://doi.org/10.1016/j.tree.2013.08.003

Morgan, M., Pages, H., Obenchain, V., & Hayden, N. (2017). *Rsamtools: binary alignment (BAM), FASTA, variant call (BCF), and tabix file import. R Package Version 1.3.0* (http://Bioconductor.Org/Packages/Release/Bioc/Html/Rsamtools.Html

Nei, M. (1973). Analysis of gene diversity in subdivided populations. *Proceedings of the National Academy of Sciences of the United States of America*, *70*, 3321–3323.

Orr, H. A., & Betancourt, A. J. (2001). Haldane's sieve and adaptation from the standing genetic variation. *Genetics*, *157*(2), 875–884.

Paradis, E., & Schliep, K. (2018). ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, *35*, 526–528.

R Core Team (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Roesti, M., Gavrilets, S., Hendry, A. P., Salzburger, W., & Berner, D. (2014). The genomic signature of parallel adaptation from shared genetic variation. *Molecular Ecology*, *23*, 3944–3956.

Roesti, M., Kueng, B., Moser, D., & Berner, D. (2015). The genomics of ecological vicariance in threespine stickleback fish. *Nature Communications*, *6*, 8767. https://doi.org/10.1038/ncomms9767; http://www.nature.com/articles/ncomms9767#supplementary-information

Roesti, M., Moser, D., & Berner, D. (2013). Recombination in the threespine stickleback genome - patterns and consequences. *Molecular Ecology*, *22*, 3014–3027.

Scheet, P., & Stephens, M. (2006). A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase. *American Journal of Human Genetics*, *78*(4), 629–644. https://doi.org/10.1086/502802

Schliep, K. (2011). phangorn: phylogenetic analysis in R. *Bioinformatics*, *27*(4), 592–593.

Schluter, D., & Conte, G. L. (2009). Genetics and ecological speciation. *Proceedings of the National Academy of Sciences of the United States of America*, *106*, 9955–9962. https://doi.org/10.1073/pnas.0901264106

Spence, R., Wootton, R. J., Barber, I., Przybylski, M., & Smith, C. (2013). Ecological causes of morphological evolution in the three-spined stickleback. *Ecology and Evolution*, *3*(6), 1717–1726. https://doi.org/10.1002/ece3.581

Stamatakis, A. (2014). RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, *30*(9), 1312–1313. https://doi.org/10.1093/bioinformatics/btu033

Terekhanova, N. V., Barmintseva, A. E., Kondrashov, A. S., Bazykin, G. A., Mugue, N. S., & Alba, M. (2019). Architecture of parallel adaptation in ten lacustrine threespine stickleback populations from the White Sea area. *Genome Biol. Evol*, *11*(9), 2605–2618. https://doi.org/10.1093/gbe/evz175

Waterston, A. R., Holden, A. V., Campbell, R. N., & Maitland, P. S. (1979). Inland waters of the Outer Hebrides. *P. Roy. Soc. Edinb. B*, *77*, 329–351.

Yeaman, S. (2015). Local adaptation by alleles of small effect. *American Naturalist*, *186*(S1), S74–S89. https://doi.org/10.1086/682405

Yeaman, S., & Whitlock, M. C. (2011). The genetic architecture of adaptation under migration-selection balance. *Evolution*, *65*(7), 1897–1911. https://doi.org/10.1111/j.1558-5646.2011.01269.x

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Haenel, Q., Guerard, L., MacColl, A. D. C., & Berner, D. (2022). The maintenance of standing genetic variation: Gene flow vs. selective neutrality in Atlantic stickleback fish. *Molecular Ecology*, 31, 811–821. https://doi.org/10.1111/mec.16269

MOLECULAR ECOLOGY

Supplemental Information for:

The maintenance of standing genetic variation: gene flow versus selective neutrality in Atlantic stickleback fish

Quiterie Haenel, Laurent Guerard, Andrew D. C. MacColl and Daniel Berner

Table of Contents:

TABLE S1 Description of the individual and pooled samples used for the study.

Sample name in study	Original sample name	Habitat	Country	Region	Locality	Latitude	Longitude	Sampling method	Read depth (pooled: median read depth per base)	DNA library type	Number of individuals pooled	Comments
BIHAR 1	BIH4 (GAD356)	Freshwater, acidic	UK (Scotland)	North Uist	ar Bhanpa			Ishmon traps	15	individual		Geographic location given in Table 1 of Magalhaes et al. 2016. Molecular Ecology
BIHAR 2	BIH25 (GAD956)	Freshwater, acidic	UK (Scotland)	North Uist	ar Bhanpa			Ishmon traps	18	individual		Geographic location given in Table 1 of Magalhaes et al. 2016. Molecular Ecology
BUAI 1	BUI5 (GAD351)	Freshwater, acidic	UK (Scotland)	North Uist	na Buaile			Ishmon traps	18	individual		Geographic location given in Table 1 of Magalhaes et al. 2016. Molecular Ecology
BUAI 2	BUI18 (GAD354)	Freshwater, acidic	UK (Scotland)	North Uist	na Buaile			Ishmon traps	24	individual		Geographic location given in Table 1 of Magalhaes et al. 2016. Molecular Ecology
FADA 1	FAD1 (GAD360)	Freshwater, acidic	UK (Scotland)	North Uist	Fadas			Ishmon traps	17	individual		Geographic location given in Table 1 of Magalhaes et al. 2016. Molecular Ecology
FADA 2	FAD13 (GAD367)	Freshwater, acidic	UK (Scotland)	North Uist	Fadas			Ishmon traps	19	individual		Geographic location given in Table 1 of Magalhaes et al. 2016. Molecular Ecology
IALA 1	IA7 (GAD368)	Freshwater, acidic	UK (Scotland)	North Uist	Ialadh			Ishmon traps	22	individual		Geographic location given in Table 1 of Magalhaes et al. 2016. Molecular Ecology
IALA 2	IA24 (GAD383)	Freshwater, acidic	UK (Scotland)	North Uist	Ialadh			Ishmon traps	20	individual		Geographic location given in Table 1 of Magalhaes et al. 2016. Molecular Ecology
SCAD 1	SC27 (GAD961)	Freshwater, acidic	UK (Scotland)	North Uist	Scadabay			Ishmon traps	19	individual		Geographic location given in Table 1 of Magalhaes et al. 2016. Molecular Ecology
SCAD 2	SC28 (GAD962)	Freshwater, acidic	UK (Scotland)	North Uist	Scadabay			Ishmon traps	16	individual		Geographic location given in Table 1 of Magalhaes et al. 2016. Molecular Ecology
Acidic pool	ANU_pool	Freshwater, acidic	UK (Scotland)	North Uist				Ishmon traps	135	pooled	100	Combined DNA from all the acidic lakes (20 individuals from each)
FEIT 1	FEI16 (GAD951)	Freshwater, basic	UK (Scotland)	North Uist	near Feithean			Ishmon traps	22	individual		Geographic location given in Table 1 of Magalhaes et al. 2016. Molecular Ecology
FEIT 2	FEI17 (GAD952)	Freshwater, basic	UK (Scotland)	North Uist	near Feithean			Ishmon traps	24	individual		Geographic location given in Table 1 of Magalhaes et al. 2016. Molecular Ecology
GROG 1	GR14 (GAD955)	Freshwater, basic	UK (Scotland)	North Uist	Grogary			Ishmon traps	24	individual		Geographic location given in Table 1 of Magalhaes et al. 2016. Molecular Ecology
GROG 2	GR18 (GAD959)	Freshwater, basic	UK (Scotland)	North Uist	Grogary			Ishmon traps	36	individual		Geographic location given in Table 1 of Magalhaes et al. 2016. Molecular Ecology
HOST 1	HOS (GAD936)	Freshwater, basic	UK (Scotland)	North Uist	Hosta			Ishmon traps	24	individual		Geographic location given in Table 1 of Magalhaes et al. 2016. Molecular Ecology
HOST 2	HOT4 (GAD817)	Freshwater, basic	UK (Scotland)	North Uist	Hosta			Ishmon traps	22	individual		Geographic location given in Table 1 of Magalhaes et al. 2016. Molecular Ecology
RIEI 1	RIE1 (GAD1149)	Freshwater, basic	UK (Scotland)	North Uist	na Rieidh			Ishmon traps	17	individual		Geographic location given in Table 1 of Magalhaes et al. 2016. Molecular Ecology
RIEI 2	RIE13 (GAD1159)	Freshwater, basic	UK (Scotland)	North Uist	na Rieidh			Ishmon traps	28	individual		Geographic location given in Table 1 of Magalhaes et al. 2016. Molecular Ecology
SAND 1	SA8 (GAD778)	Freshwater, basic	UK (Scotland)	North Uist	Sandary			Ishmon traps	16	individual		Geographic location given in Table 1 of Magalhaes et al. 2016. Molecular Ecology
SAND 2	SA9 (GAD779)	Freshwater, basic	UK (Scotland)	North Uist	Sandary			Ishmon traps	17	individual		Geographic location given in Table 1 of Magalhaes et al. 2016. Molecular Ecology
Basic pool	BUNU_pool	Freshwater, basic	UK (Scotland)	North Uist				Ishmon traps	66	pooled	100	Combined DNA from all the basic lakes (20 individuals from each)
ARDH 1	ARDH_1	Marine	UK (Scotland)	North Atlantic	North Uist, tidal lagoon	57°34'44.4"N	7°24'57.6"W	Ishmon traps	4	individual		Sequenced individually for Ihaenni et al. 2019. Evolution Letters. Raw data available from NCBI Sequence Read Archive under BioProject number PRJNA488117
ARDH 2	ARDH_2	Marine	UK (Scotland)	North Atlantic	North Uist, tidal lagoon	57°34'44.4"N	7°24'57.6"W	Ishmon traps	5	individual		Sequenced individually for Ihaenni et al. 2019. Evolution Letters. Raw data available from NCBI Sequence Read Archive under BioProject number PRJNA488117
CBEB 1	CBEB1_716	Marine	UK (Scotland)	North Atlantic	North Uist, tidal lagoon	57°38'00.0"N	7°10'00.0"W	Ishmon traps	8	individual		Sequenced individually for Ihaenni et al. 2019. Evolution Letters. Raw data available from NCBI Sequence Read Archive under BioProject number PRJNA488117
NU marine pool	NHNU_pool	Marine	UK (Scotland)	North Atlantic	North Uist, tidal lagoon			Ishmon traps	139	pooled	30	For this pool, genotype data from 10 ARDH and 20 CBEB1 individuals sequenced individually for Ihaenni et al. 2019. Evolution Letters. Raw data available from NCBI Sequence Read Archive under BioProject number PRJNA488117. Reads from the underlying individuals are available from NCBI Sequence Read Archive under BioProject number PRJNA488117
BR 1	NAO-RIL-11	Marine	Ireland	North Atlantic	North Atlantic, off the Irish coast	56°11'24.0"N	8°59'24.0"W	Research vessel	11	individual		Part of the samples taken at Site 1 in Ravinet et al. 2014. Ecology and Evolution
BR 2	NAO-RIL-17	Marine	Ireland	North Atlantic	North Atlantic, off the Irish coast	56°11'24.0"N	8°59'24.0"W	Research vessel	9	individual		Part of the samples taken at Site 1 in Ravinet et al. 2014. Ecology and Evolution
BR 3	NAO-RIL-21	Marine	Ireland	North Atlantic	North Atlantic, off the Irish coast	56°11'24.0"N	8°59'24.0"W	Research vessel	10	individual		Part of the samples taken at Site 1 in Ravinet et al. 2014. Ecology and Evolution
BR 4	NAO-RIL-23	Marine	Ireland	North Atlantic	North Atlantic, off the Irish coast	56°11'24.0"N	8°59'24.0"W	Research vessel	10	individual		Part of the samples taken at Site 1 in Ravinet et al. 2014. Ecology and Evolution
BR pool	NAO-RIL_pool	Marine	Ireland	North Atlantic	North Atlantic, off the Irish coast	56°11'24.0"N	8°59'24.0"W	Research vessel	79	pooled	21	Part of the samples taken at Site 1 in Ravinet et al. 2014. Ecology and Evolution
NL 1	BWS-NL-1	Marine	The Netherlands	North Sea	Brouwersdam, Western Scheldt estuary	51°23'39.9"N	3°52'15.0"E	Ishmon traps	8	individual		
NL 3	BWS-NL-3	Marine	The Netherlands	North Sea	Brouwersdam, Western Scheldt estuary	51°23'39.9"N	3°52'15.0"E	Ishmon traps	10	individual		
NL 4	BWS-NL-20	Marine	The Netherlands	North Sea	Brouwersdam, Western Scheldt estuary	51°23'39.9"N	3°52'15.0"E	Ishmon traps	6	individual		
NL_pool	BWS-NL_pool	Marine	The Netherlands	North Sea	Brouwersdam, Western Scheldt estuary	51°23'39.9"N	3°52'15.0"E	Ishmon traps	89	pooled	28	
DE 1	SYLT-D-05	Marine	Germany	North Sea	Sylt	55°01'14.6"N	8°26'29.2"E	Ishmon traps	10	individual		
DE 2	SYLT-D-08	Marine	Germany	North Sea	Sylt	55°01'14.6"N	8°26'29.2"E	Ishmon traps	13	individual		
DE 3	SYLT-D-43	Marine	Germany	North Sea	Sylt	55°01'14.6"N	8°26'29.2"E	Ishmon traps	13	individual		
DE pool	SYLT-D_pool	Marine	Germany	North Sea	Sylt	55°01'14.6"N	8°26'29.2"E	Ishmon traps	41	pooled	26	
IS 1	LL-ICE-538	Marine	Iceland	Greenland Sea	Lonsvik Lón	66°07'06.6"N	16°39'51.4"W	Ishmon traps	10	individual		
IS 2	LL-ICE-545	Marine	Iceland	Greenland Sea	Lonsvik Lón	66°07'06.6"N	16°39'51.4"W	Ishmon traps	11	individual		
IS 3	LL-ICE-551	Marine	Iceland	Greenland Sea	Lonsvik Lón	66°07'06.6"N	16°39'51.4"W	Ishmon traps	20	individual		
IS 4	LL-ICE-810	Marine	Iceland	Greenland Sea	Lonsvik Lón	66°07'06.6"N	16°39'51.4"W	Ishmon traps	10	individual		
IS pool	LL-ICE_pool	Marine	Iceland	Greenland Sea	Lonsvik Lón	66°07'06.6"N	16°39'51.4"W	Ishmon traps	94	pooled	25	
CA 1	IV-CA-517	Marine	Canada	North Western Atlantic, Saint Lawrence Gulf	Baie-de-Isle, Varie	48°02'55.9"N	69°21'20.7"W	Ishmon traps	11	individual		
CA 2	IV-CA-518	Marine	Canada	North Western Atlantic, Saint Lawrence Gulf	Baie-de-Isle, Varie	48°02'55.9"N	69°21'20.7"W	Ishmon traps	9	individual		
CA 3	IV-CA-529	Marine	Canada	North Western Atlantic, Saint Lawrence Gulf	Baie-de-Isle, Varie	48°02'55.9"N	69°21'20.7"W	Ishmon traps	10	individual		
CA 4	IV-CA-530	Marine	Canada	North Western Atlantic, Saint Lawrence Gulf	Baie-de-Isle, Varie	48°02'55.9"N	69°21'20.7"W	Ishmon traps	21	individual		
CA pool	IV-CA_pool	Marine	Canada	North Western Atlantic, Saint Lawrence Gulf	Baie-de-Isle, Varie	48°02'55.9"N	69°21'20.7"W	Ishmon traps	72	pooled	24	

2

TABLE S2 Genome-wide median (upper-right semimatrix) and mean (lower-left semimatrix) genetic differentiation, expressed as absolute allele frequency differentiation (AFD), and as F_{ST} (Nei's 1973 estimator G_{ST}) in parentheses, for all pairwise comparisons of marine sample pools. The underlying markers are the 1.5 million poolSeq SNPs ascertained based on the comparison of the acidic versus basic sample pool (sex chromosome and unanchored scaffolds included; ascertaining SNPs directly in the specific marine pool pairs produced very similar differentiation values). For each marine sample comparison, the SNPs were filtered for coverage (total base count across the two alleles >= 50 within each pool) and MAF (>= 0.25 across the two pools combined). Note that computing differentiation for comparisons involving acidic or basic stickleback was not meaningful, as these pools represented mixes of DNA from different replicate populations within each habitat type. However, differentiation among the freshwater populations, and between each freshwater population and North Uist marine stickleback, is described in Table S2 of Haenel et al. 2019.

	NU	**IR**	**NL**	**DE**	**IS**	**CA**
NU		0.095 (0.010)	0.123 (0.016)	0.123 (0.017)	0.134 (0.019)	0.206 (0.046)
IR	0.112 (0.022)		0.132 (0.019)	0.124 (0.017)	0.139 (0.021)	0.201 (0.044)
NL	0.148 (0.039)	0.161 (0.046)		0.082 (0.007)	0.123 (0.017)	0.190 (0.039)
DE	0.146 (0.037)	0.152 (0.041)	0.099 (0.017)		0.117 (0.015)	0.164 (0.029)
IS	0.159 (0.044)	0.169 (0.050)	0.151 (0.040)	0.145 (0.037)		0.156 (0.027)
CA	0.234 (0.092)	0.238 (0.096)	0.225 (0.087)	0.204 (0.074)	0.190 (0.063)	

TABLE S3 Characterization of the 50 adaptive SNPs, ordered by chromosome and position. The alleles give the type of polymorphism at each SNP, with the first and second nucleotide representing the allele predominant in the acidic and basic stickleback pool. The AFD values indicate the allele frequency difference between these pools. At the SNPs shaded blue, the acidic allele is the minor allele (frequency < 0.5) in all six (dark) or in at least four (light) of the marine stickleback samples. At the SNPs shaded red, the acidic allele is the major allele (frequency >= 0.5) in all six (dark) or in at least four (light) of the marine samples. The three SNPs printed in bold are the ones visualized in Figure 4.

SNP	Chr	Pos	Alleles (acidic/ basic)	acidic-basic AFD	Frequency of acidic allele in acidic pool	Frequency of acidic allele in basic pool	Frequency of acidic allele in the marine samples					
							NU	IR	NL	DE	IS	CA
1	I	879'044	C/A	0.876	0.992	0.116	0.755	0.790	0.545	0.750	0.829	1.000
2	I	8'680'470	A/C	0.914	0.944	0.030	0.146	0.155	0.340	0.225	0.375	0.178
3	I	12'858'811	T/C	0.898	0.960	0.063	0.636	0.750	0.281	0.222	0.769	0.000
4	I	25'581'749	T/A	0.860	0.967	0.110	0.145	0.121	0.398	0.484	0.322	0.476
5	I	26'278'312	C/T	0.858	0.977	0.119	0.000	0.000	0.115	0.033	0.000	0.000
6	II	4'524'607	T/C	0.853	0.945	0.093	0.454	0.333	0.492	0.286	0.796	0.781
7	II	13'793'172	T/A	0.858	0.036	0.894	0.000	0.000	0.091	0.000	0.000	0.000
8	IV	12'021'899	C/T	0.947	1.000	0.053	0.993	1.000	1.000	1.000	1.000	1.000
9	IV	12'445'286	C/A	0.903	0.955	0.052	0.951	0.911	0.762	0.587	0.872	0.873
10	IV	12'604'827	A/T	0.887	0.950	0.063	0.993	1.000	0.933	0.833	0.990	0.971
11	IV	13'957'151	T/C	0.882	1.000	0.118	0.006	0.000	0.058	0.000	0.037	0.000
12	IV	19'997'604	T/C	0.924	0.955	0.031	0.088	0.042	0.667	0.700	0.406	0.000
13	IV	20'115'769	C/T	0.872	1.000	0.128	1.000	1.000	0.543	0.524	0.067	1.000
14	IV	20'348'982	G/A	0.894	0.000	0.894	0.410	0.462	0.581	0.417	0.455	1.000
15	IV	21'780'217	A/G	0.870	0.941	0.071	0.301	0.462	0.670	0.750	0.414	0.017
16	IV	26'641'810	G/A	0.924	0.974	0.050	0.017	0.000	0.216	0.205	0.326	0.000
17	IV	33'864'589	A/G	0.854	0.977	0.123	0.489	0.455	0.504	0.295	0.108	0.573
18	V	3'953'806	C/G	0.875	0.875	0.000	0.700	0.577	0.538	0.537	0.295	0.443
19	V	8'733'705	T/C	0.870	0.949	0.079	0.556	0.407	0.298	0.267	0.417	0.140
20	V	12'751'826	C/T	0.885	1.000	0.115	0.131	0.271	0.424	0.207	0.143	0.717
21	VII	13'825'503	C/T	0.887	0.887	0.000	0.306	0.186	0.271	0.069	0.524	0.051
22	VII	19'986'348	A/G	0.960	0.960	0.000	0.326	0.278	0.310	0.132	0.084	0.000
23	VII	23'508'447	C/T	0.859	0.859	0.000	0.026	0.015	0.146	0.289	0.301	0.375
24	VIII	1'103'291	T/G	0.936	0.935	0.000	0.712	0.643	0.000	0.000	0.000	0.057
25	VIII	7'206'638	T/C	0.868	0.056	0.923	0.029	0.000	0.125	0.000	0.385	0.000
26	IX	7'878'428	G/T	0.879	0.918	0.040	0.117	0.039	0.213	0.095	1.000	0.000
27	IX	12'612'422	G/A	0.872	0.897	0.025	0.115	0.082	0.451	0.317	0.714	0.345
28	IX	13'200'359	T/C	0.936	0.989	0.053	0.988	1.000	0.842	0.909	0.930	1.000
29	IX	13'354'585	C/A	0.867	0.867	0.000	0.022	0.000	0.210	0.136	0.160	0.000
30	X	4'259'221	G/A	0.946	0.961	0.015	0.048	0.022	0.036	0.000	0.802	0.968
31	X	10'097'654	T/A	0.877	0.890	0.013	0.096	0.040	0.247	0.300	0.038	0.099
32	X	10'555'758	G/T	0.952	0.952	0.000	0.277	0.400	0.269	0.487	0.344	0.195
33	X	15'307'031	C/T	0.852	0.948	0.097	0.234	0.136	0.360	0.394	0.371	0.351
34	XI	6'489'914	T/C	0.878	0.946	0.067	0.026	0.038	0.023	0.000	0.080	0.000
35	XII	5'123'854	G/T	0.901	0.919	0.018	0.484	0.296	0.600	0.542	0.425	0.274
36	XII	6'754'066	C/A	0.864	0.864	0.000	0.000	0.000	0.167	0.000	0.167	0.240
37	XII	17'793'998	T/G	0.929	1.000	0.071	0.500	0.714	1.000	1.000	0.417	0.783
38	XIII	4'342'266	T/A	0.938	1.000	0.063	0.571	0.856	0.173	0.387	0.561	0.103
39	XIII	14'003'722	T/C	0.854	0.995	0.141	0.268	0.527	0.560	0.673	0.442	0.382
40	XIII	18'701'974	A/G	0.865	0.923	0.058	0.263	0.152	0.128	0.000	0.240	0.373
41	XIV	7'260'047	A/G	0.884	0.920	0.036	0.318	0.317	0.652	0.400	0.523	0.322
42	XV	5'954'154	T/A	0.912	0.961	0.048	0.596	0.484	0.862	0.643	0.500	0.585
43	XV	16'209'873	A/G	0.957	0.957	0.000	0.024	0.000	0.314	0.250	0.000	0.000
44	XVI	4'797'943	A/C	0.866	0.866	0.000	0.075	0.073	0.000	0.043	0.120	0.484
45	XVI	6'828'403	G/T	0.881	0.915	0.034	0.407	0.267	0.291	0.380	0.453	0.014
46	XVII	6'781'353	C/A	0.914	0.984	0.067	0.493	0.288	0.648	0.538	0.734	0.608
47	XX	7'959'446	T/A	0.852	0.063	0.915	0.195	0.088	0.556	0.714	0.375	0.636
48	XX	8'597'535	C/T	0.851	0.879	0.027	0.179	0.394	0.084	0.065	0.341	0.034
49	XX	10'599'415	A/G	0.872	0.901	0.029	0.000	0.012	0.074	0.026	0.483	0.000
50	XXI	16'133'430	C/A	0.941	0.962	0.021	0.286	0.290	0.480	0.343	0.457	0.955

TABLE S4 Compilation of the genes located within a 100 kb window around each of the 50 adaptive SNPs, ordered by chromosome and position.

SNP	Chromosome, Position	Gene ID	Gene name
1	chrI 879044	ENSGACG00000004922	ENSGACG00000004922
		ENSGACG00000004927	ENSGACG00000004927
		ENSGACG00000004929	triap1
		ENSGACG00000004934	supt5h
		ENSGACG00000004963	cox7a1
		ENSGACG00000004964	nf1b
		ENSGACG00000004992	SMCO4
2	chrI 868047	ENSGACG00000009072	grik4
3	chrI 12858811	ENSGACG00000011223	ENSGACG00000011223
		ENSGACG00000011230	ca4b
		ENSGACG00000022214	ENSGACG00000022214
		ENSGACG00000011249	ENSGACG00000011249
		ENSGACG00000011251	ywhag2
		ENSGACG00000011259	camkk1b
		ENSGACG00000011266	ENSGACG00000011266
		ENSGACG00000011279	ENSGACG00000011279
		ENSGACG00000011281	c2cd3
		ENSGACG00000011291	p4ha3
		ENSGACG00000011312	or129-1
		ENSGACG00000011316	diabloa
		ENSGACG00000011318	ENSGACG00000011318
		ENSGACG00000011320	ENSGACG00000011320
4	chrI 25581749	ENSGACG00000014600	zgc:172122
		ENSGACG00000014601	ENSGACG00000014601
		ENSGACG00000014605	u2af1
		ENSGACG00000014627	cbsa
		ENSGACG00000014641	ENSGACG00000014641
		ENSGACG00000014643	ENSGACG00000014643
		ENSGACG00000014645	stoml2
		ENSGACG00000014669	CYP4F8
5	chrI 26278312	ENSGACG00000014299	spega
		ENSGACG00000014313	CTDSP1
		ENSGACG00000021310	MIR26B
		ENSGACG00000014321	ENSGACG00000014321
		ENSGACG00000014323	obsl1a
		ENSGACG00000014324	atp1a1a.2
6	chrII 4524607	ENSGACG00000014582	pepd
7	chrII 13793172	ENSGACG00000014321	ENSGACG00000014321
		ENSGACG00000016060	ddx21
		ENSGACG00000016061	casc1
		ENSGACG00000016064	ENSGACG00000016064
		ENSGACG00000016067	mpc1
		ENSGACG00000016068	kifbp
		ENSGACG00000016070	ENSGACG00000016070
		ENSGACG00000016072	vps26a
		ENSGACG00000016077	supv3l1
		ENSGACG00000016082	hkdc1

5

8	chrIV 12021899	ENSGACG00000018224	zdhhc15b
		ENSGACG00000018229	uprt
		ENSGACG00000018231	abcb7
9	chrIV 12445286	ENSGACG00000018271	rhogd
		ENSGACG00000018273	ogt.1
		ENSGACG00000022698	ENSGACG00000022698
		ENSGACG00000018279	gcna
		ENSGACG00000018281	cetn2
		ENSGACG00000018285	nsdhl
		ENSGACG00000018286	ENSGACG00000018286
		ENSGACG00000018287	ENSGACG00000018287
		ENSGACG00000018289	fut11
		ENSGACG00000018291	rab9b
		ENSGACG00000018292	plp1a
10	chrIV 12604827	ENSGACG00000018296	nlgn3a
11	chrIV 13957151	ENSGACG00000018422	cpeb4a
		ENSGACG00000018432	stc2a
		ENSGACG00000018433	nkx2.5
		ENSGACG00000018435	bnip1a
		ENSGACG00000018438	atp6v0e1
		ENSGACG00000018439	rpl26
		ENSGACG00000018440	ppp2r2ca
12	chrIV 19997604	ENSGACG00000019554	atoh8
		ENSGACG00000019555	tmem129
		ENSGACG00000019558	rnf103
		ENSGACG00000019560	abhd18
		ENSGACG00000019563	igbp1
		ENSGACG00000019568	magt1
		ENSGACG00000019572	fbxo38
13	chrIV 20115769	ENSGACG00000019538	zgc:113425
		ENSGACG00000019540	rab33ba
		ENSGACG00000019542	hspa9
		ENSGACG00000019553	slitrk2
14	chrIV 20348982	ENSGACG00000019519	slc38a4
		ENSGACG00000019520	slc38a2
		ENSGACG00000019521	ENSGACG00000019521
		ENSGACG00000019522	arid2
		ENSGACG00000022878	ENSGACG00000022878
15	chrIV 21780217	ENSGACG00000019344	mkln1
		ENSGACG00000019341	nfyba
		ENSGACG00000019342	ENSGACG00000019342
		ENSGACG00000019343	ENSGACG00000019343
		ENSGACG00000021429	ENSGACG00000021429
		ENSGACG00000022316	MIR29A
16	chrIV 26641810	ENSGACG00000018964	ENSGACG00000018964
		ENSGACG00000018957	CDPF1
		ENSGACG00000018958	pparaa
		ENSGACG00000022207	ENSGACG00000022207
		ENSGACG00000021259	MIRLET7A3
		ENSGACG00000018960	atxn10
17	chrIV 33864589	ENSGACG00000000636	ncaph2

6

		ENSGACG00000000639	sco2
		ENSGACG00000000640	ENSGACG00000000640
		ENSGACG00000000642	flncb
		ENSGACG00000000650	kcnd2
		ENSGACG00000000652	tspan12
		ENSGACG00000000654	ing3
		ENSGACG00000000655	ENSGACG00000000655
		ENSGACG00000000657	wnt16
18	chrV 3953806	ENSGACG00000005496	usp22
		ENSGACG00000005506	cops3
		ENSGACG00000005546	NT5M
		ENSGACG00000005572	rasd1
		ENSGACG00000005578	pemt
19	chrV 8733705	ENSGACG00000002892	ENSGACG00000002892
		ENSGACG00000002871	myoz1b
		ENSGACG00000002878	synpo2lb
		ENSGACG00000002883	sec24c
		ENSGACG00000002888	ENSGACG00000002888
		ENSGACG00000002890	ENSGACG00000002890
		ENSGACG00000002901	si:ch73-127m5.1
		ENSGACG00000002906	ENSGACG00000002906
20	chrV 12751826	ENSGACG00000007789	ENSGACG00000007789
		ENSGACG00000007794	si:dkey-32n7.7
		ENSGACG00000007797	vkorc1
		ENSGACG00000007803	mapk7
		ENSGACG00000007820	ENSGACG00000007820
		ENSGACG00000007835	pax10
		ENSGACG00000007839	mmp25b
		ENSGACG00000007849	ca15b
		ENSGACG00000007888	ENSGACG00000007888
		ENSGACG00000007890	ENSGACG00000007890
		ENSGACG00000007899	ENSGACG00000007899
		ENSGACG00000007901	hrc
		ENSGACG00000007920	si:dkey-94f20.4
21	chrVII 13825503	ENSGACG00000020116	ENSGACG00000020116
		ENSGACG00000020117	HSPA8
		ENSGACG00000021746	SNORD14
		ENSGACG00000020118	ENSGACG00000020118
		ENSGACG00000020119	ENSGACG00000020119
		ENSGACG00000020120	bsx
		ENSGACG00000020121	lim2.1
22	chrVII 19986348	ENSGACG00000020345	ENSGACG00000020345
		ENSGACG00000020344	PLS3
		ENSGACG00000020346	ccdc61
		ENSGACG00000020347	itpkca
		ENSGACG00000020348	ENSGACG00000020348
		ENSGACG00000020349	ppm1nb
		ENSGACG00000020350	rtn2b
		ENSGACG00000020351	nectin3b
		ENSGACG00000020352	or133-3
		ENSGACG00000020353	ppme1
		ENSGACG00000020354	ucp2
		ENSGACG00000020355	dnajb13
		ENSGACG00000020356	rab6a

7

23	chrVII 23508447	ENSGACG00000020590	mtnr1bb
		ENSGACG00000020591	timm10b
		ENSGACG00000020592	ENSGACG00000020592
		ENSGACG00000020593	smpd1
		ENSGACG00000020594	tpte
		ENSGACG00000020595	vps36
24	chrVIII 1103291	ENSGACG00000002165	ENSGACG00000002165
		ENSGACG00000020968	5S_rRNA
		ENSGACG00000021717	5S_rRNA
		ENSGACG00000021183	U1
		ENSGACG00000021620	U5
		ENSGACG00000002169	ENSGACG00000002169
		ENSGACG00000002171	nxpe3
25	chrVIII 7206638	ENSGACG00000006030	ttc14
		ENSGACG00000006044	ENSGACG00000006044
		ENSGACG00000006048	ENSGACG00000006048
		ENSGACG00000006052	fxr1
		ENSGACG00000006076	dnajc19
26	chrIX 7878428	ENSGACG00000018741	gde1
		ENSGACG00000018744	si:dkey-44g23.5
		ENSGACG00000018746	ENSGACG00000018746
		ENSGACG00000018747	crebbpa
		ENSGACG00000018754	adcy9
27	chrIX 12612422	ENSGACG00000018022	ndst3
		ENSGACG00000018024	ugt8
28	chrIX 13200359	NULL	NULL
29	chrIX 13354585	ENSGACG00000017879	dnaja1
		ENSGACG00000017887	aptx
		ENSGACG00000017889	ENSGACG00000017889
		ENSGACG00000017892	sparcl1
		ENSGACG00000017898	odam
		ENSGACG00000017900	cnga1a
		ENSGACG00000017903	TACR3
30	chrX 4259221	ENSGACG00000002539	atp9b
		ENSGACG00000002474	tekt2
		ENSGACG00000002481	usf2
		ENSGACG00000002484	ENSGACG00000002484
		ENSGACG00000002488	naxe
		ENSGACG00000002503	scn1bb
		ENSGACG00000002516	zbtb22a
		ENSGACG00000002525	ENSGACG00000002525
		ENSGACG00000002526	galr1a
		ENSGACG00000002533	sall3b
31	chrX 10097654	ENSGACG00000005401	PTDSS1
		ENSGACG00000005407	mterf3
		ENSGACG00000005419	uqcrb
		ENSGACG00000005427	lrrc14b
		ENSGACG00000005433	gatad1
		ENSGACG00000005443	ENSGACG00000005443
		ENSGACG00000005445	fbxl2
		ENSGACG00000005492	ENSGACG00000005492
		ENSGACG00000005497	clasp2

8

		ENSGACG00000005520	ubp1
32	chrX 10555758	ENSGACG00000005895	gabbr2
		ENSGACG00000005879	galnt12
		ENSGACG00000005894	ENSGACG00000005894
33	chrX 15307031	ENSGACG00000008826	ENSGACG00000008826
		ENSGACG00000008831	ENSGACG00000008831
		ENSGACG00000008840	ENSGACG00000008840
		ENSGACG00000008842	ENSGACG00000008842
		ENSGACG00000008844	ENSGACG00000008844
		ENSGACG00000022224	ENSGACG00000022224
		ENSGACG00000021332	ENSGACG00000021332
34	chrXI 6489914	ENSGACG00000008462	tubg1
		ENSGACG00000008473	si:ch211-18i17.2
		ENSGACG00000008483	cntnap1
		ENSGACG00000008492	ezh1
		ENSGACG00000008501	ramp2
		ENSGACG00000008510	ENSGACG00000008510
		ENSGACG00000008514	ENSGACG00000008514
		ENSGACG00000008517	c1ql3b
		ENSGACG00000008519	ccdc43
		ENSGACG00000008523	fzd2
		ENSGACG00000008527	mylk5
		ENSGACG00000008532	si:ch73-141c7.1
		ENSGACG00000008535	hsd17b1
		ENSGACG00000008544	zgc:153952
		ENSGACG00000008553	atp6v0a1a
		ENSGACG00000008605	PTRF
		ENSGACG00000008607	stat3
		ENSGACG00000008634	stat5a
		ENSGACG00000008641	si:ch211-210g13.5
		ENSGACG00000008648	kcnh4a
35	chrXII 5123854	ENSGACG00000012081	prickle3
		ENSGACG00000022195	MIR124-3
		ENSGACG00000012014	ythdf1
		ENSGACG00000012016	ENSGACG00000012016
		ENSGACG00000012020	si:dkey-70p6.1
		ENSGACG00000012022	uckl1a
		ENSGACG00000012046	samd10a
		ENSGACG00000012050	emilin3a
		ENSGACG00000012054	opn7d
		ENSGACG00000012057	snpha
		ENSGACG00000012061	ENSGACG00000012061
		ENSGACG00000012076	fam110a
36	chrXII 6754066	ENSGACG00000011007	plxna2
		ENSGACG00000011016	ENSGACG00000011016
		ENSGACG00000022307	ENSGACG00000022307
		ENSGACG00000021505	ENSGACG00000021505
		ENSGACG00000021506	ENSGACG00000021506
37	chrXII 17793998	ENSGACG00000004001	tns2a
		ENSGACG00000004023	tmem106c
		ENSGACG00000004040	ENSGACG00000004040
		ENSGACG00000004044	sars1
		ENSGACG00000004085	ENSGACG00000004085
		ENSGACG00000004087	fam50a

9

		ENSGACG00000004113	ENSGACG00000004113
		ENSGACG00000004115	ENSGACG00000004115
		ENSGACG00000004120	ENSGACG00000004120
		ENSGACG00000004125	zgc:103759
		ENSGACG00000004127	gss
38	chrXIII 4342266	ENSGACG00000005923	ENSGACG00000005923
		ENSGACG00000005927	msh3
39	chrXIII 14003722	ENSGACG00000012169	si:dkeyp-14d3.1
		ENSGACG00000012181	slc15a4
		ENSGACG00000012194	glt1d1
40	chrXIII 18701974	ENSGACG00000014606	cnnm4b
		ENSGACG00000014609	ENSGACG00000014609
		ENSGACG00000014611	ENSGACG00000014611
		ENSGACG00000014617	ENSGACG00000014617
		ENSGACG00000014618	h2az2a
		ENSGACG00000014626	hs3st1l2
41	chrXIV 7260047	ENSGACG00000017091	bmp1b
		ENSGACG00000017088	antxr1b
		ENSGACG00000017093	trim69
		ENSGACG00000017101	kcnip3b
		ENSGACG00000017107	ENSGACG00000017107
		ENSGACG00000017108	prnpb
		ENSGACG00000017111	pptc7b
		ENSGACG00000017119	aplnrb
		ENSGACG00000017120	tbc1d10ab
		ENSGACG00000017126	nfkbil1
		ENSGACG00000017132	atp6v0a2a
		ENSGACG00000017144	osbp2
42	chrXV 5954154	ENSGACG00000008383	lrrc9
		ENSGACG00000008300	kif15
		ENSGACG00000008322	tdrd9
		ENSGACG00000008360	ENSGACG00000008360
		ENSGACG00000008396	pcnx4
		ENSGACG00000008408	dhrs7
		ENSGACG00000008424	ppm1aa
43	chrXV 16209873	ENSGACG00000013067	ak7b
		ENSGACG00000013078	ENSGACG00000013078
		ENSGACG00000013081	vrk1
44	chrXVI 4797943	ENSGACG00000002252	ENSGACG00000002252
		ENSGACG00000002254	ENSGACG00000002254
		ENSGACG00000002255	ENSGACG00000002255
		ENSGACG00000002255	ENSGACG00000002255
		ENSGACG00000002259	ENSGACG00000002259
45	chrXVI 6828403	ENSGACG00000002800	il1rapl1a
46	chrXVII 6781353	ENSGACG00000007358	syn2b
		ENSGACG00000007365	timp4.1
		ENSGACG00000007391	ENSGACG00000007391
		ENSGACG00000007398	FOXJ3
		ENSGACG00000007405	ppcs
		ENSGACG00000007417	utp3
		ENSGACG00000007429	ENSGACG00000007429

10

		ENSGACG00000007430	ENSGACG00000007430
		ENSGACG00000007437	si:dkey-264d12.4
47	chrXX 7959446	ENSGACG00000009832	cd4-1
		ENSGACG00000009781	si:ch211-154o6.3
		ENSGACG00000009790	ENSGACG00000009790
		ENSGACG00000009793	cops7a
		ENSGACG00000009825	si:ch73-86n18.1
		ENSGACG00000009840	usp5
		ENSGACG00000009861	p3h3
		ENSGACG00000009868	gnb3a
		ENSGACG00000009904	ENSGACG00000009904
		ENSGACG00000009907	pex5
		ENSGACG00000009910	clstn3
48	chrXX 8597535	ENSGACG00000009299	abcb4
		ENSGACG00000009330	rundc3b
		ENSGACG00000009361	cnfn
		ENSGACG00000009364	tlr21
		ENSGACG00000009368	pafah1b3
		ENSGACG00000009407	ENSGACG00000009407
		ENSGACG00000009412	ENSGACG00000009412
		ENSGACG00000009423	ENSGACG00000009423
49	chrXX 10599415	ENSGACG00000007546	si:ch73-380l3.2
		ENSGACG00000007557	ENSGACG00000007557
		ENSGACG00000007563	si:dkey-238d18.4
		ENSGACG00000007569	hsc70
		ENSGACG00000007594	ctrl
		ENSGACG00000007597	si:ch211-137j23.7
		ENSGACG00000007600	lin37
		ENSGACG00000007618	ENSGACG00000007618
		ENSGACG00000007622	hspb6
		ENSGACG00000007626	psenen
		ENSGACG00000007639	kmt2bb
		ENSGACG00000007659	igflr1
		ENSGACG00000007664	zbtb32
		ENSGACG00000007668	aplp1
50	chrXXI 16133430	ENSGACG00000000588	cdh12a
		ENSGACG00000000594	CDH10
		ENSGACG00000000591	ENSGACG00000000591
		ENSGACG00000000593	ENSGACG00000000593

11

FIGURE S1

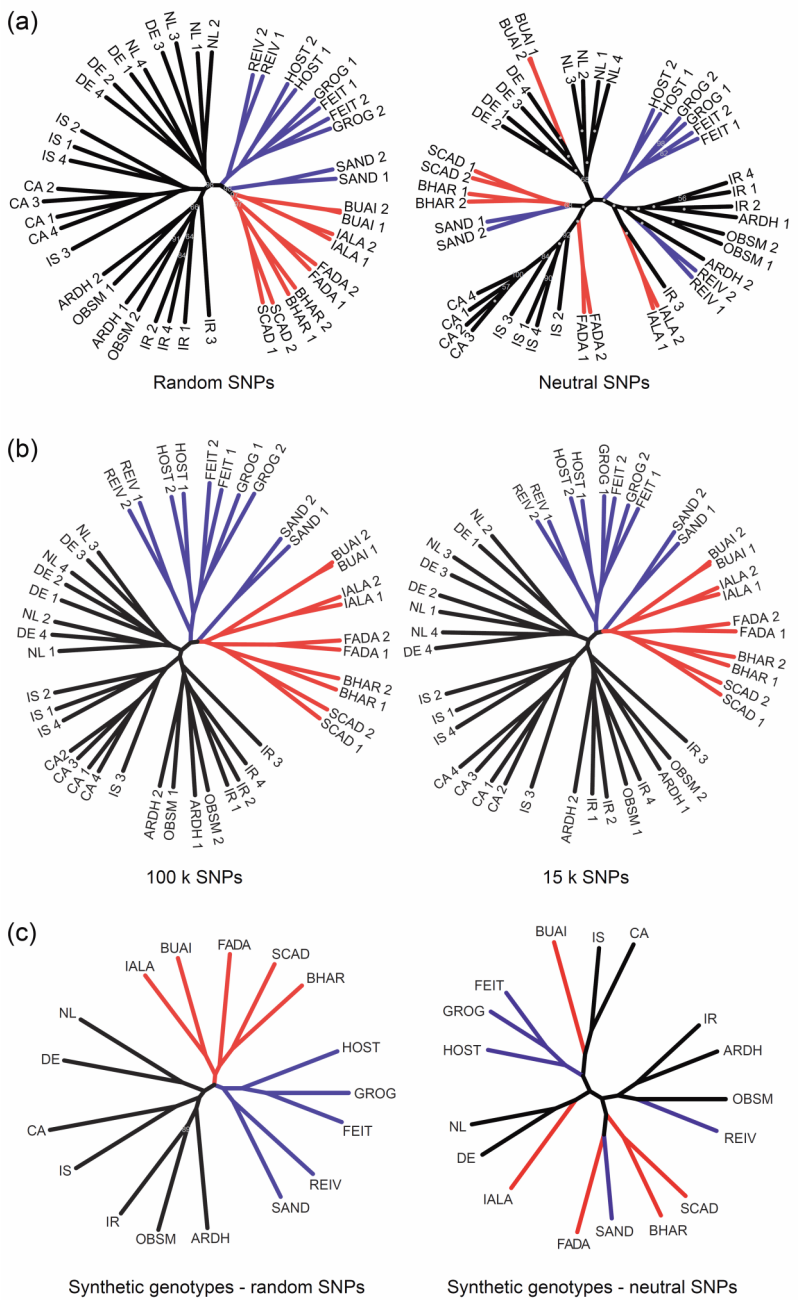

(a)
Random SNPs
Neutral SNPs
(b)
100 k SNPs
15 k SNPs
(c)
Synthetic genotypes - random SNPs
Synthetic genotypes - neutral SNPs

FIGURE S1 (a) Genealogies as presented in Figure 1c, but with bootstrap support based on 500 re-sampling iterations given by gray labels. Unlabeled nodes have 100% support, and nodes marked with a bullet point have support below 50%. (b) Robustness of the observed genealogies, as illustrated by the consistency of the tree topology despite smaller numbers of markers (100,000 and 15,000 SNPs, drawn independently from the full SNP panel). This check is presented for the random SNPs only, but similar topological robustness was observed for the neutral trees. In (c), genealogies are based on synthetic genotypes, as opposed to genotypes from true individuals. The synthetic genotypes were constructed by pooling the individual-level sequence data within each sample site, and drawing a single random nucleotide per site at each SNP. The marker resolution is similar to the one underlying the main genealogies based on individual-level data (random SNPs: n = 200,000, neutral SNPs: n = 120,170).

FIGURE S2

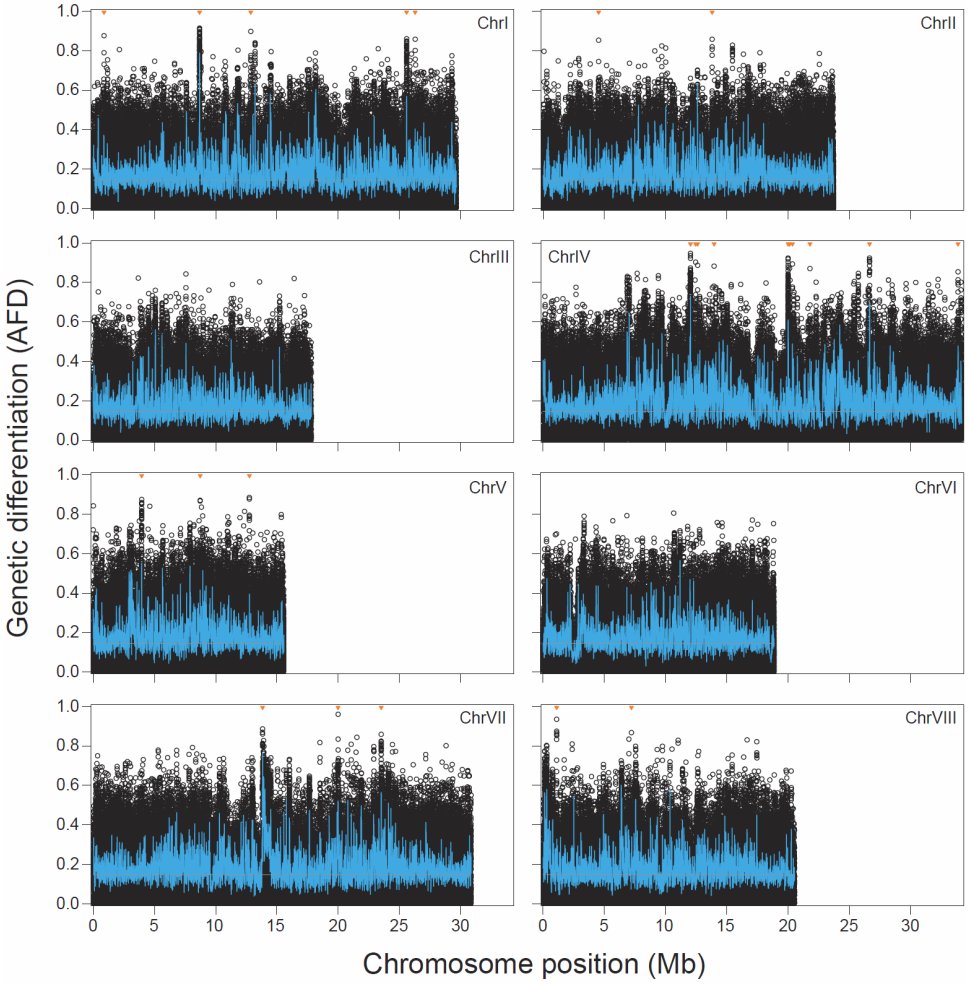

ChrI
ChrII
ChrIII
ChrIV
ChrV
ChrVI
ChrVII
ChrVIII
Genetic differentiation (AFD)
Chromosome position (Mb)
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
25
30

FIGURE S2

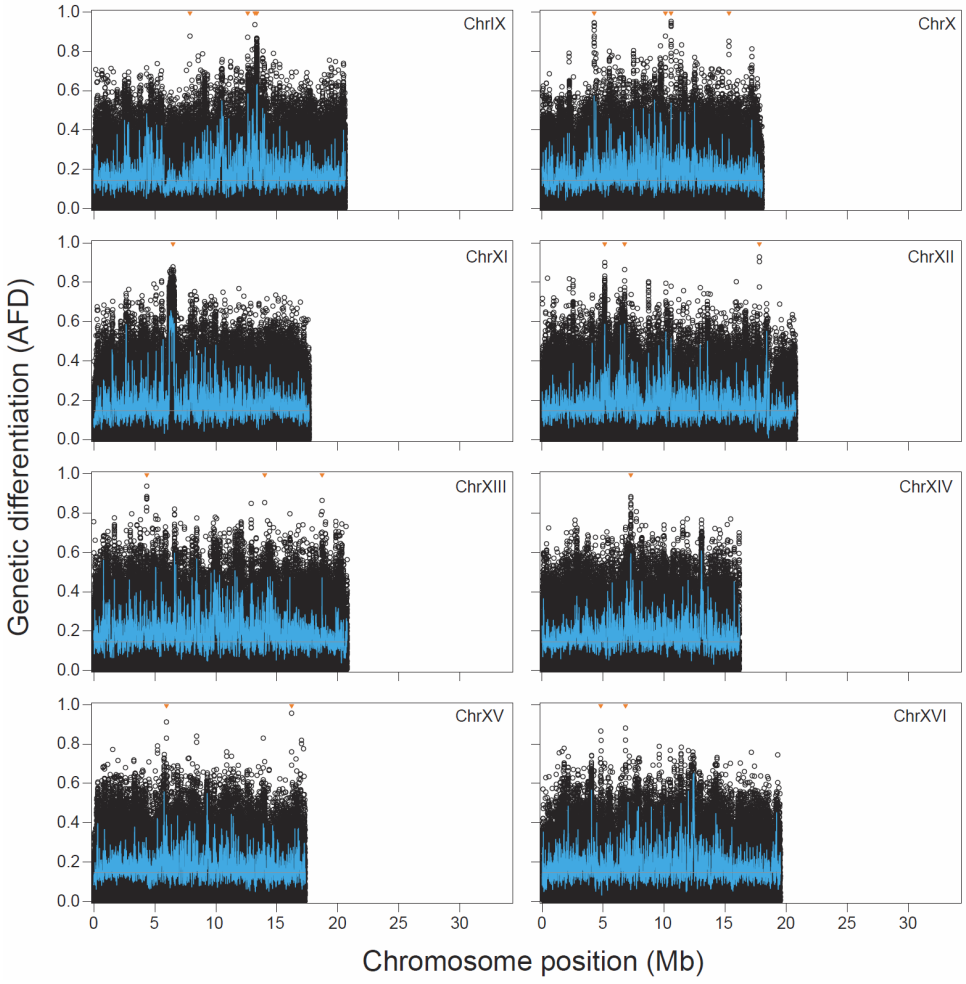

ChrIX
ChrX
ChrXI
ChrXII
ChrXIII
ChrXIV
ChrXV
ChrXVI
Genetic differentiation (AFD)
Chromosome position (Mb)
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
25
30

FIGURE S2

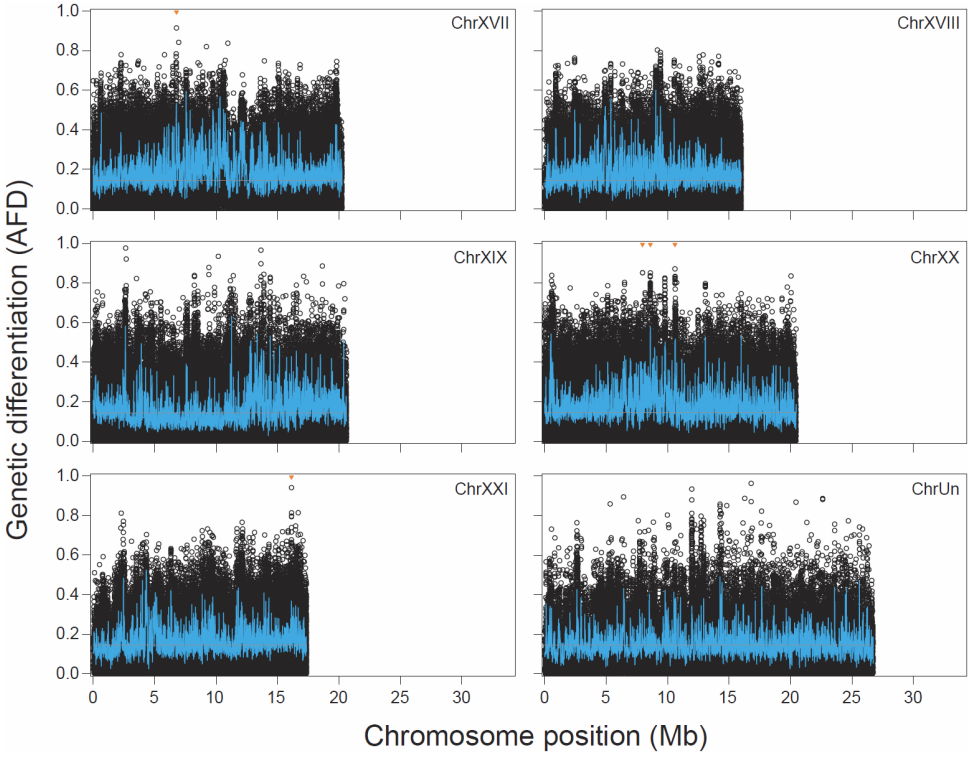


FIGURE S2 Acidic-basic differentiation across the stickleback genome. Shown is the absolute allele frequency difference (AFD) between the acidic and basic sequence pool at individual SNPs (black circles) across all stickleback chromosomes. The blue curves visualize average differentiation across sliding windows of 10 kb with 5 kb overlap (windows with fewer than six SNPs were discarded). The gray horizontal line represents the genome-wide median differentiation (0.145). The orange triangles denote the adaptive SNPs, that is, the markers exhibiting extremely strong and consistent acidic-basic differentiation that were used to explore adaptive standing genetic variation in the marine stickleback samples. Note that the sex chromosome (chrXIX) and the collection of unanchored scaffolds (chrUn) were ignored when selecting the panel of adaptive SNPs.

FIGURE S3

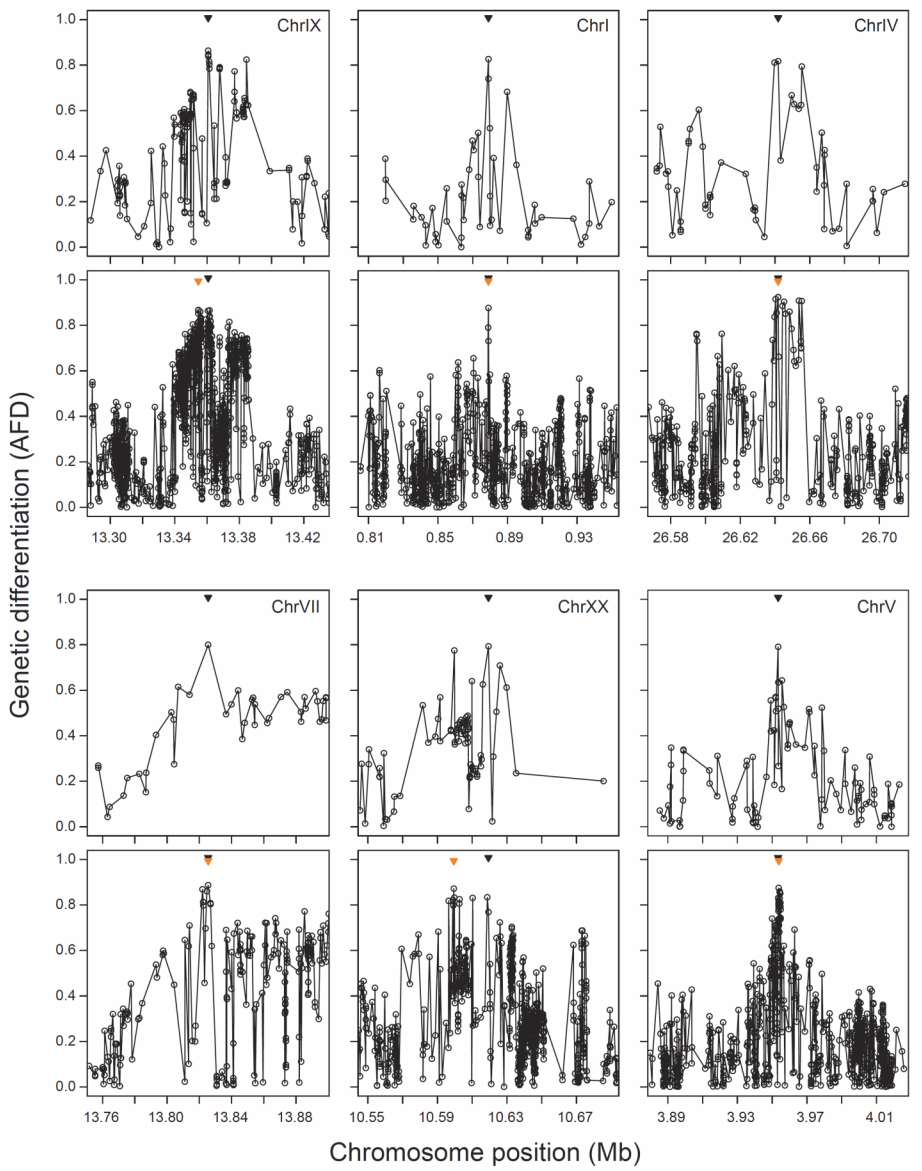

ChrIX
ChrI
ChrIV
ChrVII
ChrXX
ChrV
Genetic differentiation (AFD)
Chromosome position (Mb)

FIGURE S3

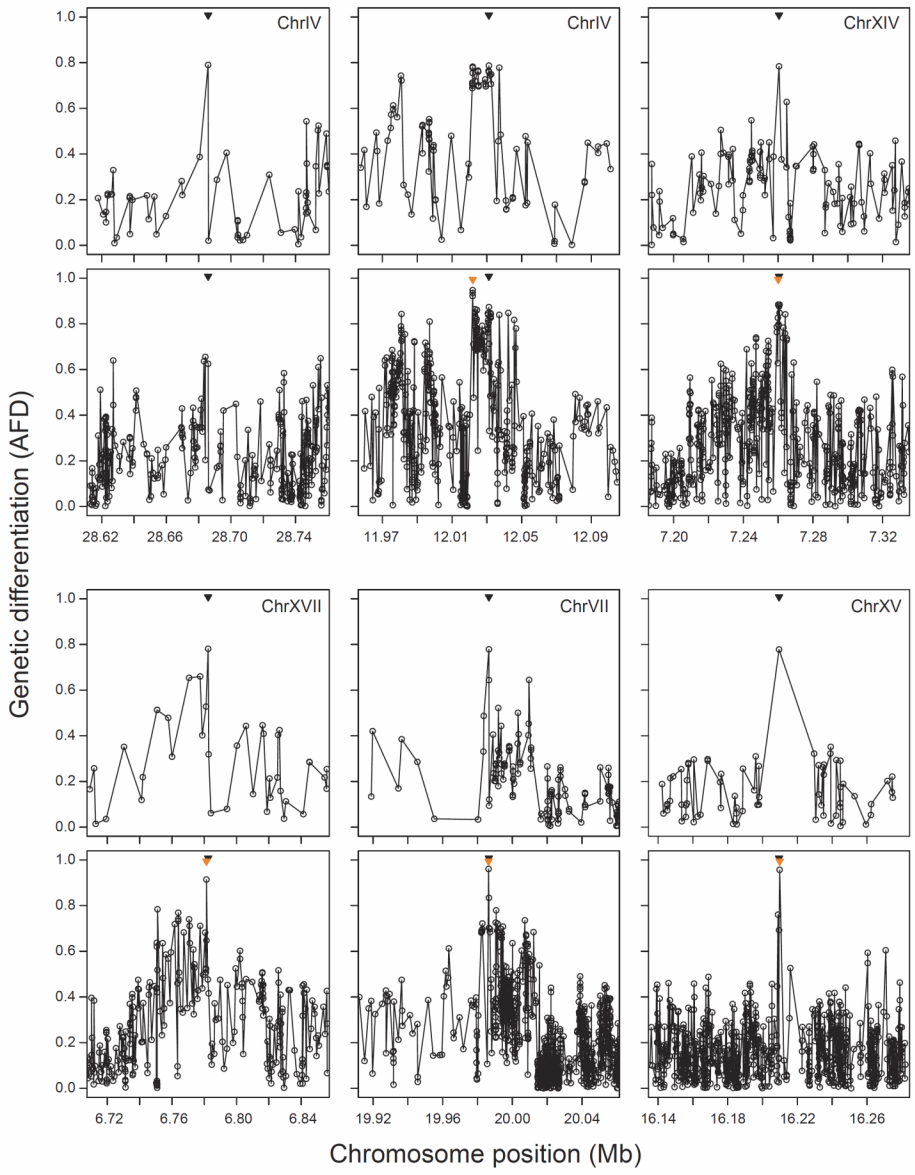

ChrIV
ChrIV
ChrXIV
ChrXVII
ChrVII
ChrXV
Genetic differentiation (AFD)
Chromosome position (Mb)

FIGURE S3

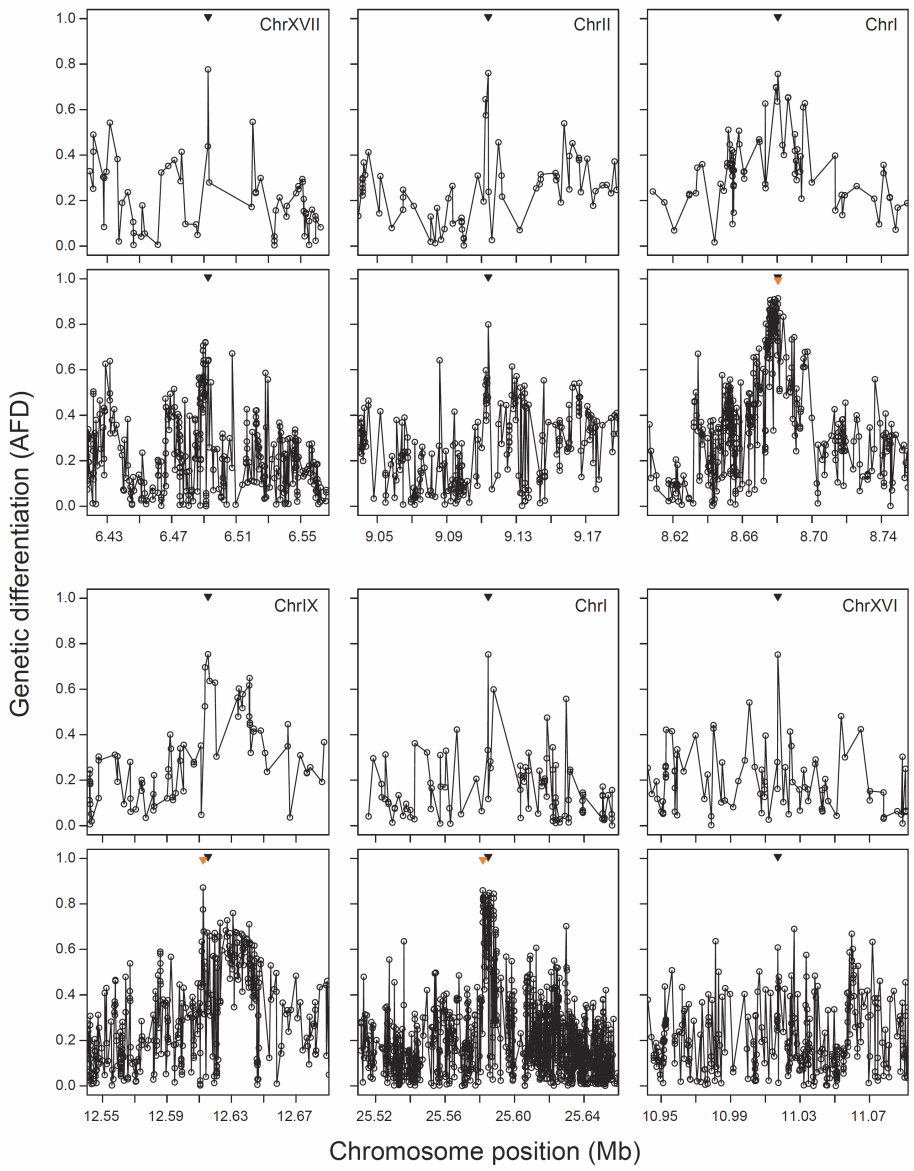


FIGURE S3 Consistency of genome regions identified as important to acidic-basic differentiation between the Haenel et al. 2019 study based on RAD sequencing, and the present work based on whole-genome sequencing. The top panels, labeled with chromosome numbers, present 18 genome regions containing a 'top core SNP' from Haenel et al. 2019, defined as markers displaying an average AFD value of >0.75 across the global acidic-basic comparison in that study.

The panels below show the same chromosome segments, but the data points are from the pooled acidic-basic comparison in the present paper. The top core SNPs from Haenel et al. 2019 are indicated by black triangles in all panels, and when present, the adaptive SNPs from the current study (selection criterion: AFD >=0.85) are added as orange triangles. All chromosome segments are 150 kb wide and centered on the top core SNPs from Haenel et al. 2019, and all tick intervals are 20 kb. Note that Haenel et al. 2019 reports 19 total top core SNPs, but one of them represents the inversion on chromosome XI (also yielding an adaptive SNP in the present study). Because of its large physical size (around 400 kb), this region is not included in this graphic but is visualized comprehensively in Figure 3 of Haenel et al. 2019, and in Figure 4 of the present study. The comparison between the two studies reveals high robustness in the identification of genome regions important to acidic adaptation, despite differences in the analyses used, and partly different panels of underlying individuals. In particular, all top core SNPs from Haenel et al. 2019 represent chromosome segments also showing strong acidic-basic differentiation in the present study. Moreover, most of the top core SNPs from Haenel et al. 2019 reside very close to (median: 424 bp), and sometimes coincide perfectly with, an adaptive SNP from the current study (maximum mismatch observed: 19.9 kb, chromosome XX).

FIGURE S4

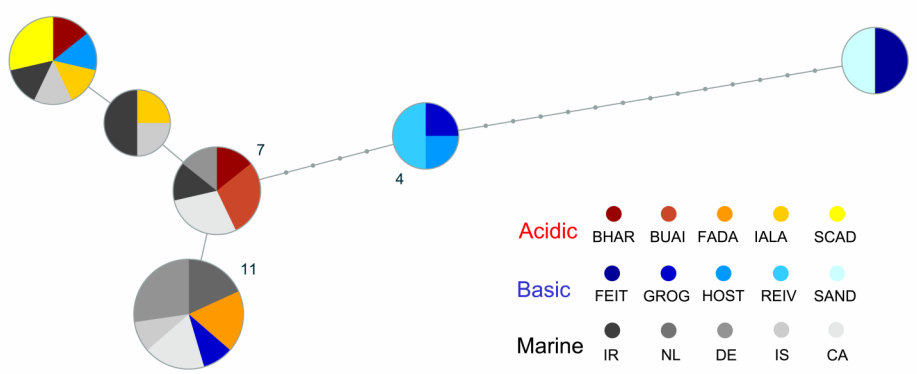


FIGURE S4 Haplotype genealogy based on polymorphisms around an exemplary adaptive SNP (SNP 8 on chromosome IV, Table S3) at which the acidic allele occurs at high frequency (near fixation) in all marine samples. All graphing conventions follow Figure 4b. Note that contrary to most adaptive SNPs, this genome region shows extensive haplotype sharing between acidic and marine stickleback, whereas basic fish harbor distinct haplotypes.

FIGURE S5

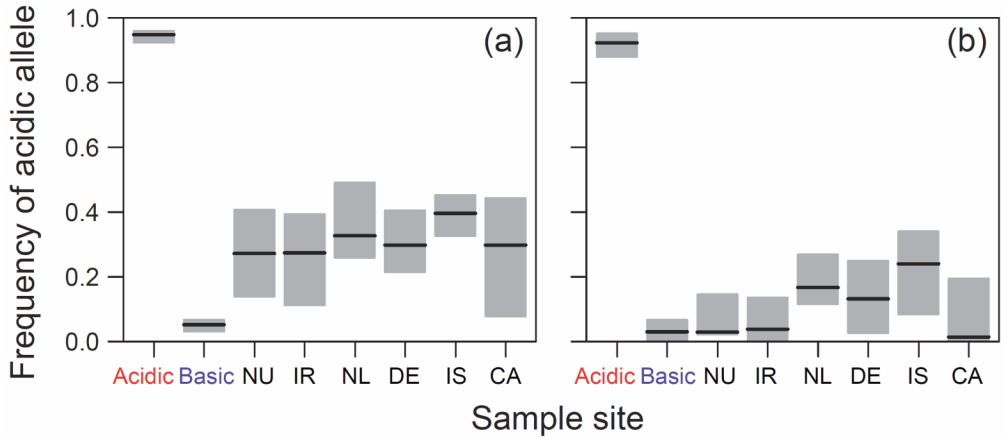


FIGURE S5 Median frequency (black lines) of the acidic alleles in all freshwater and Atlantic marine stickleback sample pools, and the associated 95% compatibility intervals (gray bars) calculated by bootstrap re-sampling allele frequencies within each sample 10,000 times. Analogously to Figure 5, panel (a) is based on all 50 adaptive SNPs, while (b) considers only the 21 adaptive SNPs at which the acidic allele was the minor allele in all marine pools. Note that for both classes of SNPs, the median frequency of the acidic alleles is relatively similar among the marine samples. In particular, the marine sample from North Uist (NU) shows absolutely no indication of elevated allele frequencies compared to the other marine samples.

Chapter 3

Clinal genomic analysis reveals strong reproductive isolation across a steep habitat transition in stickleback fish

Haenel et al. 2021, Nature Communications

nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-021-25039-y OPEN

Clinal genomic analysis reveals strong reproductive isolation across a steep habitat transition in stickleback fish

Quiterie Haenel[1✉], Krista B. Oke[2,3], Telma G. Laurentino[1], Andrew P. Hendry[3] & Daniel Berner[1✉]

How ecological divergence causes strong reproductive isolation between populations in close geographic contact remains poorly understood at the genomic level. We here study this question in a stickleback fish population pair adapted to contiguous, ecologically different lake and stream habitats. Clinal whole-genome sequence data reveal numerous genome regions (nearly) fixed for alternative alleles over a distance of just a few hundred meters. This strong polygenic adaptive divergence must constitute a genome-wide barrier to gene flow because a steep cline in allele frequencies is observed across the entire genome, and because the cline center closely matches the habitat transition. Simulations confirm that such strong divergence can be maintained by polygenic selection despite high dispersal and small per-locus selection coefficients. Finally, comparing samples from near the habitat transition before and after an unusual ecological perturbation demonstrates the fragility of the balance between gene flow and selection. Overall, our study highlights the efficacy of divergent selection in maintaining reproductive isolation without physical isolation, and the analytical power of studying speciation at a fine eco-geographic and genomic scale.

[1] Department of Environmental Sciences, Zoology, University of Basel, Basel, Switzerland. [2] College of Fisheries and Ocean Sciences, University of Alaska Fairbanks, Juneau, AK, USA. [3] Redpath Museum and Department of Biology, McGill University, Montréal, QC, Canada. ✉email: quiterie.haenel@unibas.ch; daniel.berner@unibas.ch

Deciphering the origin of species requires understanding the nature of reproductive isolation between diverging populations[1–4]. At the genomic level, reproductive isolation is typically studied through the marker-based comparison of populations that diverged relatively recently into ecologically different habitats[5–8]. Genome regions showing exceptionally strong population differentiation are inferred to harbor loci important to adaptive divergence that potentially also restrict the exchange of genetic material across larger chromosome segments, or the genome as a whole[9]. This common analytical approach generally pays insufficient attention to the fine-scale eco-geography of speciation; our understanding of the genomics of reproductive isolation can benefit greatly from investigating diverging populations across their contact zones at a high geographic resolution[10–15]. One reason is that such a clinal focus can reveal over what distance gene flow between populations occurs. Moreover, if marker resolution is sufficiently high – ideally whole-genome resolution, we can learn to what extent gene flow differs among genome regions. These details are crucial for evaluating the strength and genetic architecture of reproductive isolation. Another benefit is that fine-grained clinal analyses facilitate recognizing a possible link between reproductive isolation and ecological transitions, and hence the role of divergent adaptation in speciation[4,16].

Nevertheless, research combining a clinal perspective with the analytical power of whole-genome sequence data is largely lacking[13]. Here, we present such an investigation in a threespine stickleback fish (*Gasterosteus aculeatus*) population pair residing in parapatry (that is, contiguously) in Misty Lake and its inlet stream (Vancouver Island, British Columbia, Canada)[17,18] (Fig. 1a). This system is relatively young (postglacial, < 10,000 generations old; a generation is 1−2 years) and the populations exhibit no obvious genomic incompatibility when crossed[17–19]. Ecological differences between the lake and stream habitat, however, have driven dramatic genetically-based adaptive divergence in several traits including body shape, breeding coloration, trophic morphology and behavior[17–22] (Figs. 1b and 2a and Supplementary Fig. 1). Low-resolution marker data further indicate that phenotypic divergence between the two populations coincides with substantial genetic differentiation (F_{ST} ~0.12[23–25]). This divergence has almost certainly arisen in the face of gene flow, as opposed to reflecting secondary contact after evolution in isolation. The reason is that temporally stable spatial associations of concurrent phenotypic and genetic discontinuities of different magnitudes with lake-stream habitat transitions are ubiquitous in stickleback[26–29]. By contrast, the conditions for the evolution of neighboring lake and stream populations in initial physical isolation must be rare in general, and appear entirely implausible in watersheds within which divergent lake and stream populations have evolved repeatedly in a spatio-temporal sequence[30]. The small scale of the Misty system in particular makes the initial physical isolation of the lake and inlet stream habitat appear improbable hydro-geographically. Given the absence of physical dispersal barriers and the potential of stickleback to disperse over hundreds of meters in a few days (even against water current[31,32]), potent reproductive barriers between the lake and stream population must exist.

To characterize reproductive isolation between Misty Lake and stream stickleback at the genomic level, we perform a clinal investigation based on pooled whole-genome sequencing at a fine geographic scale. Combined with individual-based simulations tailored to this system, our study offers a fine-grained illustration of how divergent natural selection can drive and maintain reproductive isolation between populations in direct contact.

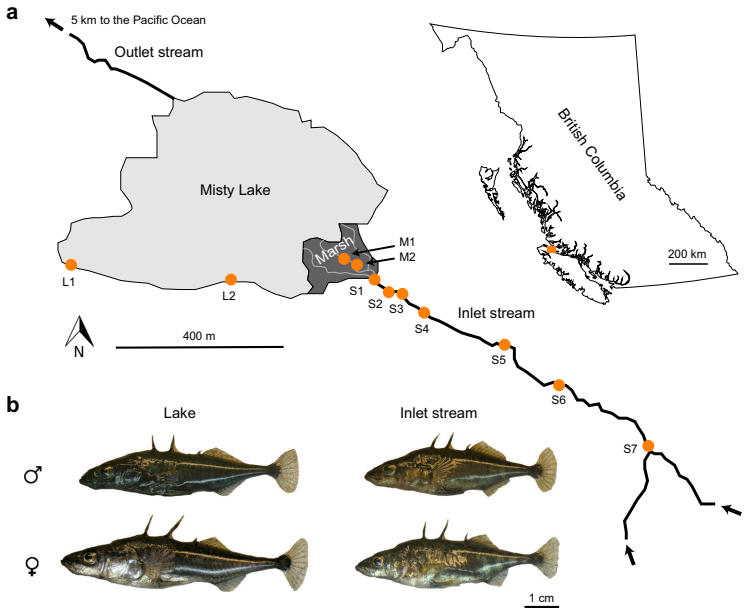


Fig. 1 The Misty Lake and inlet stream stickleback system. a Geographic situation of Misty Lake, its inlet stream, and the marsh located between these habitats (map created by the authors based on data from Google Earth). Sample sites along the lake-stream transition are indicated by orange dots (GPS coordinates and sample sizes given in Supplementary Table 1). **b** Representative lake and stream stickleback individuals of both sexes (Photo credit: Katja Räsänen).

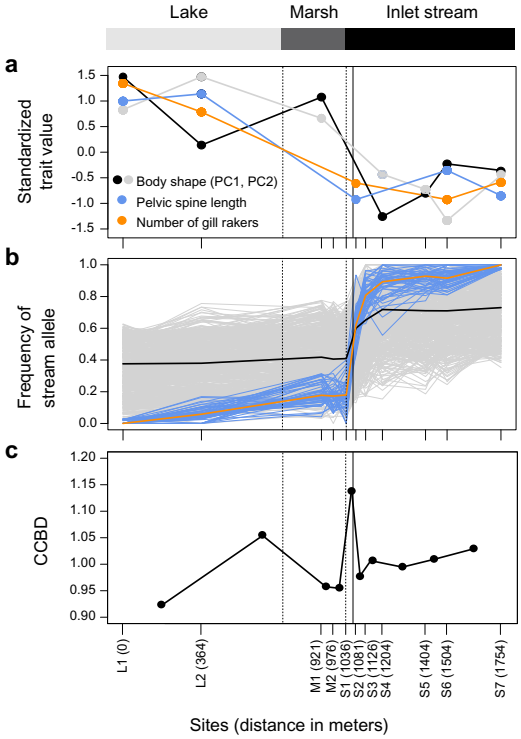


Fig. 2 Phenotypic and genomic differentiation across the lake-stream transition. a Morphology, including geometric morphometric body shape (principal component scores, details in Supplementary Fig. 1), a predator defense trait (pelvic spine length), and a foraging trait (gill raker number). For ease of presentation, only site means are shown. Data were available for a subset of the study sites only. **b** Frequency of the stream allele at the 50 independent SNPs exhibiting strong differentiation (AFD ≥ 0.97) between the most distant sites L1 and S7 (i.e. the selected SNPs, blue lines; their median frequency is shown in orange), and at 500 neutral SNPs (gray lines; median frequency in black). Source data are provided as a Source Data file. **c** Chromosome center-biased differentiation (CCBD), calculated for all pairs of neighboring samples. The midpoint between the paired samples is used as location along the gradient. In all panels, the dotted vertical lines indicate the lake-marsh and marsh-stream habitat boundaries and the black vertical line indicates the cline center estimated for the selected and neutral SNPs (1071 m). The distances on the *X*-axis represent the approximate swimming distance between each site and the lake site L1.

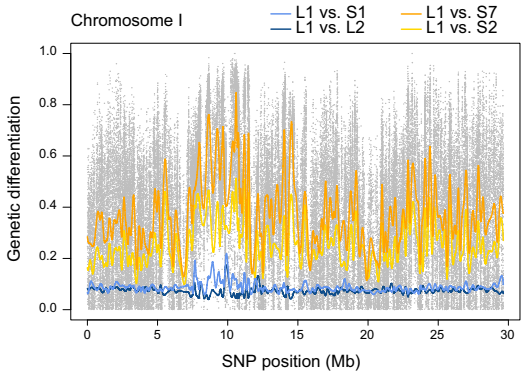


Fig. 3 Pairwise genetic differentiation between clinal sites along a representative chromosome. Genetic differentiation is expressed as absolute allele frequency differentiation (AFD) for four site comparisons, each involving the lake site L1. The colored lines represent differentiation smoothed across SNPs by non-parametric regression. For the comparison between the endpoint samples (L1-S7), differentiation is additionally shown for all SNPs individually (gray dots). Note that the differentiation profile from the L1-S1 comparison resembles the within-lake (L1-L2) comparison, except for a few genome regions particularly strongly differentiated in the L1-S1 comparison. The latter also prove exceptionally strongly differentiated between the lake and the more distant stream sites.

Results and discussion

Polygenic adaptive divergence between lake and stream stickleback. To initiate our investigation, we sampled ~56 stickleback individuals from each of 11 sites across the Misty Lake and inlet stream transition, with a distance of < 2 km between the most distant sites (L1 and S7, Fig. 1a and Supplementary Table 1). Each sample was subjected to pooled whole-genome sequencing to about 100x read depth, yielding ~1.9 million single-nucleotide polymorphisms (SNPs) as genetic markers.

We hypothesized that phenotypic and genetic divergence between the lake and stream fish maintained at a small geographic scale must be tightly linked to divergent selection between the habitats. Our first objective was therefore to identify genome regions likely targeted by selection by performing a standard pairwise population comparison. We here focused on our two most distant lake and stream sites (L1 and S7, Fig. 1a), assuming that these represented our stickleback samples the least influenced by gene flow, and hence the best adapted to local lake or stream ecology. For this site pair, we quantified the magnitude of genetic differentiation, expressed as the absolute allele frequency difference AFD[33], across all genome-wide SNPs.

This genome scan revealed a median differentiation of 0.35 across all SNPs (expressed as F_{ST}: 0.14), thus confirming the strong overall genomic differentiation between the lake and stream population indicated previously by sparser marker data[18,23–25]. However, numerous genome regions exhibited much stronger differentiation between the two most distant samples, and dozens of SNPs proved fixed for alternative alleles (differentiation along a representative chromosome is presented in Fig. 3, and along all chromosomes in Supplementary Fig. 2). We next defined all autosomal SNPs exhibiting AFD ≥ 0.97 in this site comparison as exceptionally highly differentiated ($n = 162$, ~0.01 percent of all autosomal SNPs; genome-wide AFD and F_{ST} distributions are visualized in Supplementary Fig. 3). These SNPs were considered to tag distinct high-differentiation genome regions only if they were separated by at least 50 kb. From each of the independent SNP clusters defined in this way, we then chose one representative high-differentiation SNP at random, yielding our panel of 50 selected SNPs representing putative genomic targets of strong divergent selection. Interrogating the gene annotation of the stickleback genome revealed that only one out of the 50 selected SNPs lay within a coding gene sequence, consistent with primarily regulatory evolution during stickleback diversification[34]. (The same result was obtained when expanding this check to the full 162 high-differentiation SNPs: only seven mapped to coding sequences of four different genes.) Overall, our comparison of a single site pair makes clear that genomic divergence in Misty Lake-stream stickleback is strong and highly

polygenic; divergent selection likely targets hundreds of loci, as inferred in other lake-stream stickleback systems[25,35,36].

Strong reproductive isolation at the habitat transition. Having obtained genomic evidence of the general presence of divergent selection, we took advantage of our full spatial set of samples to explore how tightly divergence was related to eco-geography. Graphing the frequency of the stream allele (i.e., the major allele at the site S7) at the selected SNPs across all 11 clinal sites uncovered a dramatic genetic shift over a physical stream distance of merely 45 meters, starting at the transition of the marsh to the stream (blue and orange lines in Fig. 2b; the stream site S1 is located in immediate proximity to this transition). This habitat transition also coincided roughly with a major shift in ecologically important traits, as revealed by morphological data from this and previous studies for a subset of our sites (Fig. 2a; note that the spatial resolution of the phenotypic data around the marsh-stream transition is low). To expand our focus beyond adaptive genetic and phenotypic variation, we next delimited two categories of SNPs unlikely to be physically close to loci under divergent selection, hence reflecting genome-wide background differentiation by drift. These included 500 SNPs chosen at random from all genome-wide autosomal SNPs deviating by no more than 0.1% from the genome-wide median AFD in the L1-S7 comparison (hereafter called our neutral SNPs), and a SNP panel derived analogously based on just half the median AFD (0.175; our loDiff SNPs). Inspecting the frequency of the stream allele (here defined as the allele more frequent in sample S7 than L1) at the neutral and loDiff SNPs revealed genetic clines as sharp as in the selected SNPs, again starting at the marsh-stream transition (gray and black lines in Fig. 2b and Supplementary Fig. 4). To compare the location of the genetic clines among our marker categories more formally, we fitted classical geographic cline models[37] to the stream allele frequencies for each of the 50 selected SNPs, and for a random subset of 50 markers from the neutral and loDiff SNP panels. This indicated similar cline centers among all SNP categories, located at ~1070 m, consistent with our visual inference (Fig. 4 left, Supplementary Fig. 5).

Together, these spatially fine-grained analyses reveal a remarkably tight association between ecology, phenotype, and genome-wide genetic variation. Divergent selection thus not only maintains adaptive divergence, but simultaneously generates strong barriers to gene flow across the entire genome. What are these barriers? A major contribution to reproductive isolation must arise directly from local adaptation, in the form of reduced performance of migrants and hybrids between the habitats[1,2,4,38,39]. Field transplant experiments in a European lake-stream stickleback system indicate that these reproductive barriers alone can reduce lake-stream gene flow by ~80%[40]. Given that phenotypic differentiation in this European lake-stream system is weak compared to other lake-stream systems[41], selection against migrants and hybrids should cause an even stronger reproductive barrier in the phenotypically highly divergent Misty stickleback. Adaptive divergence may additionally entail some sexual isolation, although experiments in Misty[42] and European[43] lake-stream stickleback indicate that this barrier is relatively weak. Moreover, reproductive isolation across the lake-stream habitat transition is plausibly promoted by habitat preference[31]. However, in no parapatric lake-stream system assessed so far have crossing experiments under standardized laboratory conditions[17–19,36,44] indicated intrinsic inferiority of hybrids (F1, F2) or backcrosses. Our evolutionarily young study system thus offers no support for the idea that steep genetic clines and the underlying strong reproductive isolation reflect primarily intrinsic, ecology-independent genetic incompatibilities having become spatially coupled with clines at loci under (potentially weak) divergent ecological selection[45,46].

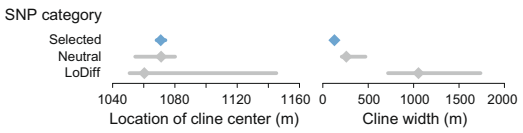


Fig. 4 Cline parameter estimates for different SNP categories. Estimation of the geographic location of the cline center and the width of the cline by cline modeling for the selected SNPs (blue; AFD ≥ 0.97 in the L1-S7 comparison), and the neutral (AFD near the genome-wide median) and loDiff (AFD near half the median) SNPs (gray). The diamonds indicate the median value across the 50 independent SNPs in each marker category, and the bars give the associated bootstrap 95% compatibility intervals. Source data are provided as a Source Data file.

Gene flow is restricted to a narrow zone. Having uncovered reproductive isolation at the habitat transition, we next asked how strong this isolation is. An informative pattern in our clinal genomic data was that for all three SNP categories (selected, neutral, loDiff), the frequency of the stream allele increased substantially upstream of the marsh-stream transition – but only up to sampling site S4, that is, over c. 150 m (black and gray curves in Fig. 2b and Supplementary Fig. 4). Beyond this location, allele frequencies consistently proved stable across the remaining stream section investigated. (This interpretation does not conflict with a subtle allele frequency shift between S6 and S7, especially at the selected SNPs; this is expected because SNP ascertainment was conditioned on allele frequencies from the site S7 in all three marker categories.) In agreement with visual inference, cline modeling estimated a median cline width of just 127 m for the selected SNPs, although somewhat wider clines were estimated for the neutral and especially the loDiff SNPs (Fig. 4 right and Supplementary Fig. 5). However, we suspected that the latter may represent an artifact of model fitting, which was assessed with simulated data differing exclusively in the magnitude of differentiation between the populations (Supplementary Fig. 6). This confirmed that identically abrupt allele frequency breaks lead to increasing cline width estimates and to increasing spread in cline center and cline width with decreasing AFD between the populations, as observed empirically across our SNP categories (Fig. 4). Our cline modeling thus yields no indication that allele frequency clines are wider in the neutral or loDiff SNPs than in the selected SNPs.

Collectively, our analyses show that beyond a few hundred meters upstream of the marsh-stream transition, reproductive isolation must already be so strong that a homogenizing effect of gene flow from the lake is no longer detectable in any marker category. Misty Lake and stream stickleback thus support models suggesting that ecologically based divergent selection on numerous loci may jointly drive reproductive isolation strong enough to block gene flow across the genome as a whole[1,47–49].

Although stream allele frequencies in all three SNP categories increased substantially over a few hundred meters upstream of the marsh-stream transition, these frequencies remained largely stable downstream of the transition, that is, all the way from site S1 across the marsh to the most distant lake site (Fig. 2b and Supplementary Fig. 4). This indicates asymmetry in gene flow; genetic variation flows predominantly from the lake (and marsh; this habitat will be discussed in a later section) into the lower reach of the stream than in the opposite direction. While this asymmetry may be influenced by differences in dispersal behavior between lake and stream fish, we believe that a main reason is

imbalance in relative population sizes: according to estimates from mark-recapture data, the lake population is substantially larger than the stream population[50]. Moreover, contrary to the lake, the stream is a linear habitat; a relatively smaller fraction of the total stream population may have the opportunity to disperse into the lake than vice versa[3]. The asymmetry in gene flow provides further evidence that beyond a few hundred meters upstream of the habitat transition, reproductive isolation must be strong across the whole genome. If the latter was not the case, one would expect cline centers to be displaced upstream in the neutral and especially the loDiff SNPs relative to the selected SNPs. However, our cline modeling confirms similar cline center locations among our marker categories (Fig. 4 and Supplementary Figs. 5 and 7).

Gene flow in the contact zone is heterogeneous across the genome. The occurrence of gene flow from the lake to the lower reach of the stream allowed us to ask if some genome regions were more strongly isolated by divergent selection than others. A pattern relevant to this question emerged when quantifying genetic divergence by pairwise genomic comparisons of each of the sites L2 to S7 to the lake site L1, and graphing the resulting population differentiation along chromosomes. This analysis revealed that the lowest stream site (S1; located right at the marsh-stream transition) was generally differentiated only trivially from the lake across most of the genome, hence producing chromosomal differentiation profiles closely resembling those from the within-lake (L1-L2) site comparison (compare the light-blue and dark-blue curves in Fig. 3 and Supplementary Fig. 2). Interestingly, however, a small number of genome regions exhibited substantially stronger differentiation in the L1-S1 comparison (e.g., ~10 Mb on chromosome I, Fig. 3; note that owing to the higher sensitivity of AFD to weak population differentiation compared to F_{ST}[33], this pattern was easier to discern with the former differentiation metric, Supplementary Fig. 8). This led us to hypothesize that our stream site closest to the lake (S1) was overwhelmed by gene flow from the lake, with appreciable genomic differentiation from the lake maintained in genome regions under the strongest divergent selection only. If true, these specific regions should, compared to randomly chosen genome regions, exhibit exceptionally strong differentiation between the lake and the stream population in general, including between the two samples from the endpoints of our geographic gradient (L1, S7). This prediction was confirmed by simulations estimating the magnitude of L1-S7 differentiation for the two categories of genome regions (i.e., highly differentiated in the L1-S1 comparison and randomly chosen) expected if drift was similar between these regions (Fig. 5). Genome regions particularly highly differentiated between the lake and the closest stream site (S1) thus harbor loci under strong divergent selection that partly resist homogenization by gene flow. We speculate that during the formation of the lake-stream stickleback pair, initial divergence at these loci set the stage for genomically more widespread adaptive divergence that now overall constitutes a strong reproductive barrier between the populations[1,47,49].

While this heterogeneity in gene flow between the populations concerned relatively small genome regions, we also obtained evidence of heterogeneous gene flow at the scale of whole chromosomes. Specifically, stickleback (like eukaryotes in general[51]) exhibit substantially elevated recombination rates near the chromosome peripheries compared to the chromosome centers[52]. Under this recombination rate distribution, theory predicts that polygenic divergent selection with gene flow causes relatively elevated population differentiation in chromosome

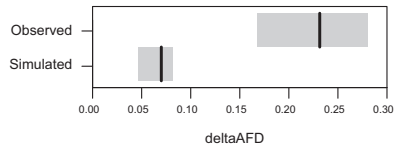


Fig. 5 High-differentiation regions maintained by divergent selection at the habitat transition. The upper row represents the empirically observed difference in AFD (deltaAFD) in the L1-S7 site comparison between high-differentiation regions and randomly chosen regions ($n = 178$ each). These regions were identified based on the L1-L2 and L1-S1 comparisons. The lower row represents deltaAFD calculated analogously for simulated high-differentiation and control loci evolving to similar baseline differentiation under drift alone ($n = 178$ each). The black vertical lines indicate median values and gray boxes represent bootstrap 95% compatibility intervals. DeltaAFD observed empirically is much greater than expected under drift alone, indicating that the high-differentiation regions identified in the L1-S1 comparison must be under particularly strong divergent selection between Misty Lake and inlet stream stickleback in general.

centers[53,54]. The reason is that in the chromosome centers, maladaptive foreign chromosome segments recombining into the locally favored genomic background will tend to be longer and hence to harbor a greater number of locally maladaptive alleles, thus making their selective elimination more efficient. We thus quantified the magnitude to which genetic differentiation is elevated across chromosome centers relative to the peripheries (i.e., chromosome center-biased differentiation, CCBD[53,54]) for all pairwise combinations of neighboring samples. This bias proved greatest for the S1-S2 sample comparison (Fig. 2c), that is, at the cline center location estimated for all SNP categories. This supports the notion of an antagonism between selection and gene flow in the lowest reach of the inlet stream.

Collectively, our analyses indicate that reproductive isolation between the Misty Lake and the inlet stream population is very strong, allowing the evolution of the two populations largely unconstrained by gene flow. Nevertheless, the lowest section of the stream represents a zone in which selection is opposed by ongoing gene flow from the lake. This gene flow must involve hybridization between lake and stream fish, not just dispersal of lake fish into the lower stream section. The reason is that both heterogeneous genomic divergence in the L1-S1 sample comparison and the CCBD peak in the S1-S2 comparison require differential lake-stream gene flow among genomic regions, hence hybridization. Nevertheless, investigating in more detail based on individual-level sequence data to what extent haplotypes typical of the lake and stream populations are mixed by recombination within the contact zone before they are eliminated by selection is an exciting future opportunity in this stickleback system.

Strong reproductive isolation in simulations of polygenic divergence. We have inferred from empirical patterns that reproductive isolation between parapatric stickleback populations reflects a by-product of adaptive divergence. To support the plausibility of our interpretation theoretically, we tailored individual-based simulations to the Misty stickleback system. We assumed nine demes in a linear array, with dispersal occurring in the beginning of every generation between contiguous demes only (stepping-stone model; Fig. 6a). Considering empirical population size estimates[50], the first (lake) deme was specified to be larger than all other (stream) demes together. The two habitats were under polygenic divergent selection, with fitness being a function of genetic variation at 100 loci. All loci were

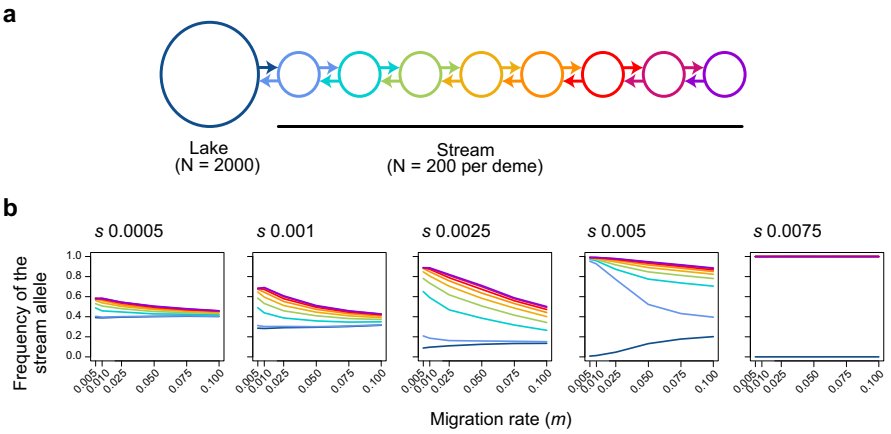


Fig. 6 Simulated divergence with gene flow across the lake-stream habitat transition. a Schematic of the stepping-stone model. Arrows indicate migration between neighboring demes. **b** Median frequency of the allele favored in the stream across 100 loci under selection after 1000 generations of evolution, averaged across 20 simulation replications for different combinations of migration rate and selection strength (selection coefficient *s* given on top of each panel). The demes are color coded as in **a**. In the rightmost panel, all stream demes are fixed for the stream allele hence the lines overlap.

polymorphic in the beginning of the simulations, thus mimicking standing genetic variation well known to underlie adaptive diversification in stickleback[29,34,55–59]. After 1000 generations of evolution, we examined clinal patterns in the frequency of the stream allele (analogously to our empirical analyses of the selected SNPs) resulting from different combinations of dispersal rates between the demes and strengths of divergent selection between the habitats.

A first observation from these simulations was that strong adaptive divergence (and hence necessarily the associated reproductive barriers) established easily across the habitat transition, even for combinations of relatively small selection coefficients and high dispersal rates (Fig. 6b). Specifically, selection coefficients just below 0.01 were already sufficient to allow complete differentiation between the lake and all stream sites across all dispersal rates considered (up to 0.1). Second, we found that whenever gene flow prevented complete lake-stream divergence, the stream site closest to the lake was particularly strongly constrained by gene flow, whereas more distant stream sites showed relatively similar allele frequencies (e.g., Fig. 6b, selection coefficients of 0.0025 or 0.005 combined with relatively high dispersal rates). This pattern closely resembled the shape of the cline in allele frequencies observed upstream of the marsh-stream transition in all SNP categories (Fig. 2b and Supplementary Fig. 4). All these observations remained qualitatively consistent across several robustness checks (Supplementary Fig. 9).

The simulations support our empirically based conclusion that adaptive divergence from abundant standing genetic variation has produced strong reproductive isolation in the absence of physical barriers. Our study thus provides insights into the genomic architecture of adaptive divergence: previous theory has emphasized that in the face of gene flow, adaptive divergence is promoted by the physical linkage of adaptive alleles, as produced by inversions[60–62]. However, in Misty Lake-stream stickleback, we found no indication of divergence in inversion polymorphisms. Importantly, the three large inversions often involved in adaptive divergence between pelagic and benthic stickleback ecotypes[29,34,55,57] proved monomorphic across our samples. Not denying the importance of chromosomal rearrangements in adaptive divergence in some study organisms, our whole-genome clinal investigation highlights that polygenic selection per se, without any particular physical arrangement of the targeted loci, can be sufficient for the emergence of strong divergence and reproductive isolation in the face of gene flow[47,49].

Perturbation of gene flow-selection balance by an unusual ecological event. Our clinal phenotypic data and allele frequencies in all SNP categories revealed that the stickleback inhabiting the marsh (sites M1 and M2) are genetically very similar to the true lake fish (L1, L2). Nevertheless, in a previous study, microsatellite loci designed to discriminate Misty Lake and inlet stream fish indicated hybridization in the marsh[22]. Similarly, the frequency of the stream allele at our selected SNPs was slightly elevated in the marsh relative to the lake samples (although this may be attributable to the ascertainment of these SNPs, see above). This raised the possibility that the marsh might allow a modest degree of genetic mixing between the lake and the stream population despite being strongly dominated by lake fish. If true, we hypothesized that changes in the level of dispersal from the lake or the stream into the marsh, as mediated by a physical perturbation of the system, should drive a measurable shift in the genetic composition of the marsh fish.

Evaluating this hypothesis was made possible by exceptionally intense rainfall during our main sampling period, causing an unusual rise in inlet stream discharge and lake water level that for a few days flooded the marsh that normally is above water level (Supplementary Fig. 10). To assess the genomic consequences of this event, we complemented our standard marsh sample (M1) taken before the flood by two additional samples from the same site, taken during the flood and 1 year later. The comparison of these temporal samples at lake-stream population-distinctive SNPs (AFD ≥ 0.75 in the lake pool [sites L1 and L2 combined] to stream pool [S6 and S7] comparison) revealed a striking decline (often to zero) of the stream allele frequency during the flood, that is, within a few days (Fig. 7). Although our pooled sequence

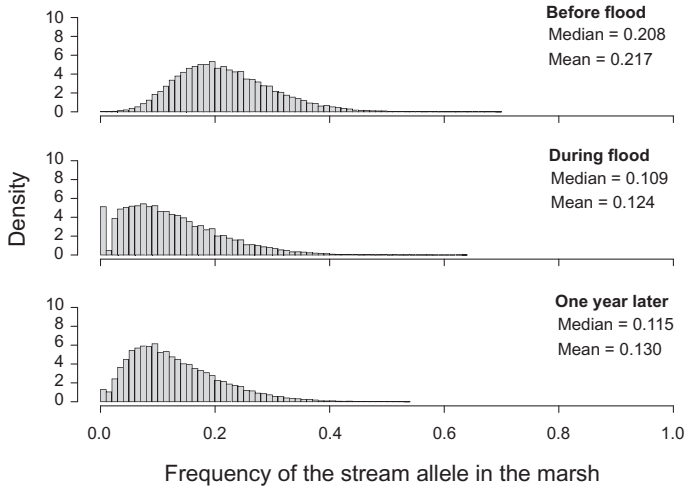


Fig. 7 Impact of a flood on the genetic composition of stickleback in the marsh. Shown is the distribution of the frequency of the stream allele across 49,677 SNPs exhibiting strong lake-stream differentiation (AFD ≥ 0.75 in the comparison between the lake pool [sites L1 and L2 combined] and the stream pool [sites S6 and S7 combined]) in stickleback sampled at the marsh site M1 at three different time points. The median and mean of the distributions are also given. Source data are provided as a Source Data file.

data precluded inspecting haplotype structure, the speed of this genetic change clearly indicates extensive dispersal of lake fish into the marsh, facilitated by easier access to the latter habitat. This conclusion is supported by simulations suggesting that at least 90% of the stickleback residing at the marsh site before the flood must have been replaced by migrants from the lake during the flood (Supplementary Fig. 11). Interestingly, this perturbation in allele frequencies at the marsh site appeared partly offset one year later (e.g., the number of SNPs monomorphic for the lake allele declined by 76%; Fig. 7), indicating selection against the lake migrants and/or the new immigration of stream individuals from the nearby contact zone.

Our genomic analysis of temporally replicated samples from the marsh supports the idea that stickleback in this habitat represent a genomic mix between the lake and stream population caused by hybridization[22]. Nevertheless, genetic material from the lake population vastly predominates in the marsh, and we demonstrate that this imbalance can become nearly complete temporarily by an unusual short-term ecological modification of dispersal opportunities. Further disentangling the relative importance of selection and gene flow as determinants of allele frequencies within this eco-geographically intermediate habitat will benefit from direct information on the local selective conditions and individual-level genomic sequence data.

To summarize, our investigation of parapatric stickleback has demonstrated a tight link between ecology, polygenic adaptive divergence, and whole-genome reproductive isolation, thereby illustrating how adaptation and speciation can be two sides of the same coin. Genetic exchange between the diverging populations, however, has not ceased completely but continues within a narrow contact zone. We show that the balance between homogenizing gene flow and divergent selection in this zone is fragile and can shift quickly when habitats are perturbed. Our work highlights the power of whole-genome sequencing at a fine spatial scale and across multiple time points to inform the eco-geography and genetic architecture of speciation.

Methods

Stickleback sampling and phenotypic analysis. Stickleback were captured with unbaited minnow traps at 11 sites in Misty Lake and its inlet stream between May and July 2016, during the breeding season (the marsh site M1 was additionally sampled in August 2017). Sample sizes ranged from 40 to 62 individuals per site (details on the locations and samples given in Supplementary Table 1). From each individual, a dorsal spine was clipped and stored in 95% ethanol for DNA extraction. All individuals were immediately released.

To allow qualitatively linking genomic to phenotypic differentiation along the geographic gradient, we performed a geometric morphometric body shape analysis. For this, 40 individuals from a subset of our sites (L1, L2, M1, S4, S5, S6, and S7) were photographed on their left side with a standard scale using a digital camera (Canon PowerShot G11, Canon, Tokyo, Japan). All photographs were digitized with tpsDIG2 (life.bio.sunysb.edu/morph/) by the same investigator (KBO) in haphazard order by placing 14 landmarks used in previous studies in the same system[24,63]. Using the geomorph R package[64,65], the resulting coordinates were aligned and generalized Procrustes analysis performed, yielding principal components of body shape variation among individuals[64,66]. In addition, we retrieved data for two ecologically important traits related to predator defense (pelvic spine length) and foraging (number of gill rakers) from another subset of our sites (L1, L2, S2, S6, and S7) studied in a previous phenotypic study of Misty Lake and inlet stickleback[67]. All these phenotypic data were mean-centered and standardized to allow visualization on the same scale. All animal work in this study was conducted in accordance with the Animal Use Protocol from McGill University.

DNA library preparation and sequencing. DNA was extracted individually from the dorsal spine clip of each of the 701 total stickleback using the Quick-DNA TM Miniprep Plus Kit (Zymo Research, Irvine, CA, USA), following the manufacturer protocol. For enzymatic tissue digestion, the spines were minced with micro spring scissors to maximize DNA yield. Following DNA quantification using a Qubit fluorometer (Invitrogen, Thermo Fisher Scientific, Wilmington, DE, USA), individuals were pooled to equal molarity without PCR-enrichment to obtain a single DNA library per sampling site (and per time point in the case of the marsh site M1). The 13 total libraries were paired-end sequenced to 151 base pairs on 10 total lanes of an Illumina HighSeq2500 instrument, producing a median read depth per base pair of 103x across the libraries (min = 51, max = 133; Supplementary Table 1). Combined with the relatively large number of individuals per site, this high read depth is expected to allow estimating allele frequencies with high precision[68].

SNP discovery. Raw sequences reads were parsed by sampling site and aligned to the third-generation assembly[69] of the 447 Mb stickleback reference genome[34] by using Novoalign (http://www.novocraft.com/products/novoalign/; settings given in the Supplementary Software). The Rsamtools R package[65,70] was then used to convert the alignments to BAM format, and to perform base counts at all genome-

wide positions for each sample using the *pileup* function. To identify informative single-nucleotide polymorphisms (SNPs), we first combined nucleotide counts from the two lake samples (L1 and L2) and from the two most upstream inlet samples (S6 and S7) into lake and stream pools. Genomic positions then qualified as SNPs for further analysis if they exhibited a read depth between 50 and 400x within each pool (to exclude poorly sequenced and repeated regions), and a minor allele frequency (MAF) of at least 0.25 across the two pools combined (to ensure a high information content[71]). Throughout the study, allele frequencies were calculated directly from raw nucleotide counts. The 1,920,596 SNPs passing these filters were genotyped in all 13 samples separately.

Quantifying clinal genomic differentiation. Genetic differentiation along the lake-stream gradient was quantified by two approaches. The first relied on the frequency of the stream allele at selected, neutral, and loDiff SNPs. The selected SNP category comprised markers showing genetic differentiation ≥ 0.97 between the geographically most distant samples (L1 and S7). Throughout the study, we quantified genetic differentiation by the absolute allele frequency difference AFD[33], although a few key analyses were repeated by using F_{ST}. We considered only nuclear markers, and ignored the sex chromosome (XIX) because it was enriched for high-differentiation SNPs relative to the autosomes, as expected from its reduced population size and hence stronger drift. Including the sex chromosome, however, always had a trivial influence on the results. If multiple high-differentiation SNPs were < 50 kb apart, they were considered a cluster from which only one SNP was chosen at random to ensure statistical independence, resulting in a panel of 50 independent selected SNPs. As a resource for future investigations, we retrieved from the reference genome annotation all genes located within a 50 kb window centered at each of the selected SNPs, producing a gene compilation (provided as Supplementary Data 1) likely containing numerous genes under divergent lake-stream selection in the Misty system. The annotation was also used to assess if the selected SNPs were located within or outside coding gene sequences. The neutral SNPs, in turn, represented 500 markers chosen at random among all autosomal SNPs for which AFD deviated from the genome-wide median differentiation in the L1-S7 comparison (c. 0.35) by no more than 0.1%, again applying a 50 kb spacing threshold. The loDiff SNPs, finally, represented markers weakly differentiated relative to the genome-wide median; they were chosen analogously to the neutral SNPs, but targeting an AFD deviation of 0.1% around just half the genome-wide median differentiation (c. 0.175). At all selected, neutral, and loDiff SNPs, we then defined the nucleotide relatively more frequent in the S7 than the L1 sample as the stream allele. Finally, the frequency of the stream alleles was calculated for each sample and visualized along the geographic gradient. In the second approach, we calculated genetic differentiation at all genome-wide SNPs for each pairwise combination of the first lake sample (L1) and all other samples, requiring a read depth between 50 and 200x within each sample. The values obtained were visualized along chromosomes, raw and/or smoothed by non-parametric regression using the *smooth.spline* R function (band width 0.1). Genome-wide median differentiation for these site comparisons, and for all comparisons between neighboring sites, is given in Supplementary Fig. 12, expressed as both AFD and F_{ST}.

Cline modeling. As a numerical complement to our visual cline analyses, we fit our allele frequency data to classical geographic cline models implemented in the HZAR R package[37]. We here used the sampling site-specific frequencies of the stream allele, the total nucleotide counts underlying these frequencies, and the geographic locations as input data, set allele frequency intervals to the observed maximum values, and assumed two independent tails (models without tails produced qualitatively similar results). We considered all 50 selected SNPs for modeling, and random subsets of the same size from the neutral and loDiff SNPs. For each SNP, cline fitting was run in 10 replicates, and the median maximum likelihood estimate of cline center and cline width across these replicates was recorded. We then compared these key cline features among the SNP categories based on the median values across SNPs and the associated 95% bootstrap compatibility intervals (10,000 resamples). The raw data distributions are provided in Supplementary Fig. 5. Because the ascertainment of the SNPs in all three categories was contingent on the magnitude of differentiation between our most distant sites (L1, S7), thus generating subtle allele frequency shifts between each of these sites and their adjacent site (most pronounced in the selected SNPs; Fig. 2b), we repeated all cline modeling by excluding the sites L1 and S7. This produced qualitatively similar results leading to the same conclusion (Supplementary Fig. 7a, b).

Our visual analysis revealed stable allele frequencies over several hundred meters in the upper reach of the stream for all three SNP categories (Fig. 2b). By contrast, cline modeling indicated an increase of cline width from the selected to the neutral and loDiff SNPs. This discrepancy led us to hypothesize that the inverse relationship between cline width and AFD among the SNP categories may be a modeling artifact, which was confirmed by simulation (details presented in Supplementary Fig. 6).

Inferring selection-gene flow antagonism from high-differentiation regions. While inspecting the comparison L1-S1, we observed that some genome regions showed remarkably strong genetic differentiation compared to the overall undifferentiated genome-wide background. This led us to speculate that in these specific regions, genetic variants particularly strongly favored in the stream were maintained at elevated frequency at the S1 site, while the remainder of the genome was overwhelmed by gene flow from the lake. Assuming particularly strong divergent selection on these genome regions, we predicted that they should exhibit exceptionally strong differentiation between the lake and the stream population in general, including in the comparison L1-S7. To evaluate this idea, we first subtracted the mean AFD value in the L1-L2 comparison (considered the differentiation baseline within the lake habitat) from the corresponding value in the L1-S1 comparison for each genome-wide 10 kb sliding window (5 kb overlap) containing at least 5 SNPs. For the most highly differentiated 0.5% of these windows (high-differentiation windows, HDW; $n = 178$; median AFD difference between the comparisons L1-S1 and L1-L2: 0.132), considered regions under strong divergent lake-stream selection, and for the same number of windows chosen at random as control (non-HDW; median AFD difference: 0.011), we then calculated mean AFD in the L1-S7 comparison.

Because the HDW by definition exhibited elevated differentiation in the L1-S1 comparison, somewhat stronger L1-S7 differentiation in these windows relative to random windows was expected even if the HDW were strongly differentiated in the former comparison just by chance. Comparing the two categories of windows thus required a benchmark, which was obtained by individual-based simulation. We here constructed n haploid individuals by first generating 178 biallelic (1, 0) loci (non-HDL) at which the stream allele (1) occurred at a frequency specified by a random draw from the uniform distribution bounded between 0.05 and 0.5 (i.e., the stream allele was always the minor allele, as observed empirically at the site S1; Fig. 2b). Another set of loci (HDL) was generated analogously, except that the frequency of the stream allele was increased by 0.1, corresponding to the observed difference in median L1-S1 AFD between the HDW and non-HDW. We then allowed this population to evolve neutrally (i.e., to drift) by drawing offspring for each new generation at random with replacement. All loci were unlinked, and random assortment of alleles was achieved by swapping alleles between the haplotypes within pairs of offspring. After g generations, we determined median AFD before vs. after evolution for the HDL and non-HDL. The combination of n and g was chosen to produce drift at the non-HDL approximating the median AFD observed across the non-HDW in the L1-S7 comparison ($n = 200$, $g = 1000$; higher values for both variables produced similar results but required longer simulation). This simulation was replicated 25 times.

Finally, we calculated the difference in median AFD in the L1-S7 comparison between the empirical HDW and non-HDW, and the median difference in AFD achieved during simulated evolution at the HDL and non-HDL across the 25 replicates, both referred to as deltaAFD. Uncertainty around these point estimates was obtained by bootstrapping windows (empirical data) and replicates (simulated data) 10,000 times. Elevated empirical relative to simulated deltaAFD would indicate that regions exhibiting the strongest differentiation in the L1-S1 comparison also show exceptionally strong differentiation in the L1-S7 comparison relative to the genome-wide baseline, consistent with these regions being targets of particularly strong divergent selection between lake and stream stickleback.

Inferring selection-gene flow antagonism from CCBD. The combination of polygenic selection, gene flow, and a reduced crossover rate in chromosome centers compared to chromosome peripheries causes relatively elevated population differentiation in chromosome centers (chromosome center-biased differentiation, CCBD[51–54]). To explore the strength of selection-gene flow antagonism along the geographic gradient, we thus quantified to what extent genetic differentiation was biased toward chromosome centers for all 10 pairwise comparisons of neighboring samples. For this, we defined the outer 5 Mb on either side of a chromosome as high-crossover rate periphery and the remainder of the chromosome as low-crossover rate center[52]. Next, we divided median AFD across all central SNPs by median AFD across the peripheral SNPs for each chromosome within each site pair. For each pair, we then treated the median across these ratios as CCBD, and graphed this metric along the lake-stream gradient by using the midpoint between the neighboring samples as geographic location. As a robustness check, this analysis was repeated by using as site pairs each of the samples L2 to S7 combined with L1, which produced very similar results supporting the same conclusion (Supplementary Fig. 13).

Individual-based simulations of divergence with gene flow. To explore theoretically how selection can drive and maintain reproductive isolation in the presence of gene flow, we conducted individual-based forward simulations using a diploid stepping-stone expansion of the model in Berner and Roesti[54]. Our standard model involved nine total demes arranged in a one-dimensional array, with adjacent demes connected by migration (Fig. 6a). The first deme ($n = 2000$) represented the (larger) lake population while all other demes (each $n = 200$) represented stream sites. In the beginning of each generation, a fraction of m individuals was chosen at random from each deme to migrate into the neighboring deme on either side (juvenile migration). A total fraction of $2m$ thus emigrated from each deme, except for the demes located at the endpoints of the array, for which this fraction was m. The migration phase was followed by selection and reproduction. We modeled polygenic divergent selection by assuming a total of 100 biallelic unlinked loci under divergent selection, with one allele favored in the first

deme and the other allele favored in all other demes. The simulations started with the frequency of both alleles set to 0.5 in all demes, to minimize the probability of the stochastic loss of adaptive variation[54]. Loci contributed additively to fitness; each maladaptive allele reduced an individual's fitness from the local fitness optimum of one by s, the selection coefficient. Individual fitness was then scaled by the mean population fitness and determined an individual's probability to be drawn for reproduction. Individuals reproduced as hermaphrodites and were allowed to mate more than once, each mating producing one offspring. Mating was repeated until the number of offspring re-established initial local deme size. The offspring cohort then replaced the parental deme (discrete generations) and entered the migration phase.

We explored a range of combinations of migration rates (0.005−0.1) and selection coefficients (0.0005−0.0075). All simulations were run for 1000 generations, a time span shown by preliminary runs over up to 7000 generations to allow approaching migration-selection balance (Supplementary Fig. 14a, b). All parameter combinations were replicated 20 times. Results were visualized by plotting for each deme at generation 1000 the mean across simulation replications of the median frequency of the allele favored in the stream across all 100 loci for all combinations of migration rates and selection coefficients.

The robustness of the standard model described above was assessed by a number of additional simulations, presented in Supplementary Fig. 9. We here considered models with a multiplicative (as opposed to additive) contribution of each locus to overall fitness; a lower number of selected loci (10); physical linkage among all loci by assuming a single chromosome undergoing uniformly distributed crossover during mating (as opposed to independent segregation); locus-specific selection coefficients drawn at random from the exponential distribution with a rate equal to $1/s$ (as opposed to identical selection coefficients among loci); and greater population size imbalance by setting the lake deme ten times (as opposed to 1.25x) larger ($n = 10{,}000$) than all stream demes combined ($n = 125$ per deme).

Quantifying the impact of a flood on stickleback in the marsh. For the marsh site (M1), three temporally replicated samples were available: before, during, and one year after a flood. The former represents the sample also used in all previous analyses; the latter two samples were processed in exactly the same way. To maximize the sensitivity for detecting gene flow, we here considered only SNPs highly differentiated (AFD ≥ 0.75) between the lake pool (sites L1 and L2 combined) and the stream pool (sites S6 and S7), and sequenced to a minimum read depth of 50x within each temporal sample. For the 49,677 SNPs thus obtained, we visually compared the frequency of the stream allele among the samples.

Because this analysis indicated massive dispersal of lake stickleback into the marsh during the flood, we explored by simulation how much of such dispersal was needed to drive the observed change in allele frequencies. For the SNPs above, we here averaged allele frequency data from the marsh before the flood and those from the nearest lake site (L2), considering a wide range of relative proportions of the latter (10%−100%). Comparing visually the resulting (mixed) allele frequency distributions to the one observed during the flood allowed a qualitative assessment of the proportion of lake dispersers into the marsh during the flood. This proportion was additionally explored using Approximate Bayesian Computation (ABC) (details given in Supplementary Fig. 11). Unless stated otherwise, all our analyses and simulations were implemented in the R language[65].

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

Data availability

All raw Illumina sequences, demultiplexed by site (and sampling period for the site M1) are available from the European Nucleotide Archive (accession numbers ERS4388731-ERS4388743) under the project PRJEB37366. Raw genome-wide nucleotide counts for all sites (and temporal replicates) are provided on Dryad (https://doi.org/10.5061/dryad.c59zw3r67). Source data are provided with this paper.

Code availability

Codes are provided as Supplementary Software.

Received: 8 June 2020; Accepted: 20 July 2021;
Published online: 11 August 2021

References

1. Rice, W. R. & Hostert, E. E. Laboratory experiments on speciation: what have we learned in 40 years? *Evolution* **47**, 1637–1653 (1993).
2. Coyne, J. A., Orr, H. A. *Speciation* (Sinauer Associates, Inc. Sunderland, 2004).
3. Gavrilets, S. *Fitness Landscapes and the Origin of Species* (Princeton University Press, 2004).
4. Sobel, J. M., Chen, G. F., Watt, L. R. & Schemske, D. W. The biology of speciation. *Evolution* **64**, 295–315 (2010).
5. Ellegren, H. et al. The genomic landscape of species divergence in Ficedula flycatchers. *Nature* **491**, 756–760 (2012).
6. Martin, S. H. et al. Genome-wide evidence for speciation with gene flow in Heliconius butterflies. *Genome Res.* **23**, 1817–1828 (2013).
7. Toews, D. P. et al. Plumage genes and little else distinguish the genomes of hybridizing warblers. *Curr. Biol.* **26**, 2313–2318 (2016).
8. Elgvin, T. O. et al. The genomic mosaicism of hybrid speciation. *Sci. Adv.* **3**, e1602996 (2017).
9. Wu, C. I. The genic view of the process of speciation. *J. Evol. Biol.* **14**, 851–865 (2001).
10. Ryan, S. F. et al. Patterns of divergence across the geographic and genomic landscape of a butterfly hybrid zone associated with a climate gradient. *Mol. Ecol.* **26**, 4725–4742 (2017).
11. Stankowski, S., Sobel, J. M. & Streisfeld, M. A. Geographic cline analysis as a tool for studying genome-wide variation: a case study of pollinator-mediated divergence in a monkeyflower. *Mol. Ecol.* **26**, 107–122 (2017).
12. Pulido-Santacruz, P., Aleixo, A. & Weir, J. T. Morphologically cryptic Amazonian bird species pairs exhibit strong postzygotic reproductive isolation. *Proc. R. Soc. B.* **285**, 20172081 (2018).
13. Rafati, N. et al. A genomic map of clinal variation across the European rabbit hybrid zone. *Mol. Ecol.* **27**, 1457–1478 (2018).
14. Westram, A. M. et al. Clines on the seashore: The genomic architecture underlying rapid divergence in the face of gene flow. *Evol. Lett.* **2**, 297–309 (2018).
15. Capblancq, T., Després, L. & Mavárez, J. Genetic, morphological and ecological variation across a sharp hybrid zone between two alpine butterfly species. *Evol. Appl.* **00**, 1–16 (2020).
16. Schilthuizen, M. Ecotone: speciation-prone. *Trends Ecol. Evol.* **15**, 130–131 (2000).
17. Lavin, P. A. & McPhail, J. D. Parapatric lake and stream sticklebacks on northern Vancouver Island: disjunct distribution or parallel evolution? *Can. J. Zool.* **71**, 11–17 (1993).
18. Hendry, A. P., Taylor, E. B. & McPhail, J. D. Adaptive divergence and the balance between selection and gene flow: lake and stream stickleback in the Misty system. *Evolution* **56**, 1199–1216 (2002).
19. Berner, D. et al. Quantitative genetic inheritance of morphological divergence in a lake-stream stickleback ecotype pair: implications for reproductive isolation. *J. Evol. Biol.* **24**, 1975–1983 (2011).
20. Sharpe, D. M. T., Räsänen, K., Berner, D. & Hendry, A. P. Genetic and environmental contributions to the morphology of lake and stream stickleback: implications for gene flow and reproductive isolation. *Evol. Ecol. Res.* **10**, 849–866 (2008).
21. Raeymaekers, J. A., Delaire, L. & Hendry, A. P. Genetically based differences in nest characteristics between lake, inlet, and hybrid threespine stickleback from the Misty system, British Columbia, Canada. *Evol. Ecol. Res.* **11**, 905–919 (2009).
22. Hanson, D., Moore, J.-S., Taylor, E. B., Barrett, R. D. & Hendry, A. P. Assessing reproductive isolation using a contact zone between parapatric lake-stream stickleback ecotypes. *J. Evol. Biol.* **29**, 2491–2501 (2016).
23. Moore, J.-S., Gow, J. L., Taylor, E. B. & Hendry, A. P. Quantifying the constraining influence of gene flow on adaptive divergence in the lake-stream threespine stickleback system. *Evolution* **61**, 2015–2026 (2007).
24. Kaeuffer, R., Peichel, C. L., Bolnick, D. I. & Hendry, A. P. Parallel and nonparallel aspects of ecological, phenotypic, and genetic divergence across replicate population pairs of lake and stream stickleback. *Evolution* **66**, 402–418 (2012).
25. Stuart, Y. E. et al. Contrasting effects of environment and genetics generate a continuum of parallel evolution. *Nat. Ecol. Evol.* **1**, 0158 (2017).
26. Berner, D., Grandchamp, A.-C. & Hendry, A. P. Variable progress toward ecological speciation in parapatry: stickleback across eight lake-stream transitions. *Evolution* **63**, 1740–1753 (2009).
27. Deagle, B. E. et al. Population genomics of parallel phenotypic evolution in stickleback across stream-lake ecological transitions. *Proc. R. Soc. B* **279**, 1277–1286 (2012).
28. Ravinet, M., Prodoehl, P. A. & Harrod, C. Parallel and nonparallel ecological, morphological and genetic divergence in lake-stream stickleback from a single catchment. *J. Evol. Biol.* **26**, 186–204 (2013).
29. Roesti, M., Kueng, B., Moser, D. & Berner, D. The genomics of ecological vicariance in threespine stickleback fish. *Nat. Commun.* **6**, 8767 (2015).
30. Caldera, E. J. & Bolnick, D. I. Effects of colonization history and landscape structure on genetic variation within and among threespine stickleback (Gasterosteus aculeatus) populations in a single watershed. *Evol. Ecol. Res.* **10**, 575–598 (2008).
31. Bolnick, D. I. et al. Phenotype-dependent native habitat preference facilitates divergence between parapatric lake and stream stickleback. *Evolution* **63**, 2004–2016 (2009).

32. Moore, J.-S. & Hendry, A. P. Can gene flow have negative demographic consequences? Mixed evidence from stream threespine stickleback. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **364**, 1533–1542 (2009).
33. Berner, D. Allele Frequency Difference AFD - an intuitive alternative to FST for quantifying genetic population differentiation. Genes **10**, 308 (2019).
34. Jones, F. C. et al. The genomic basis of adaptive evolution in threespine sticklebacks. *Nature* **484**, 55–61 (2012).
35. Feulner, P. G. et al. Genomics of divergence along a continuum of parapatric population differentiation. *PLoS Genet.* **11**, e1004966 (2015).
36. Laurentino, T. G. et al. Genomic release-recapture experiment in the wild reveals within-generation polygenic selection in stickleback fish. *Nat. Commun.* **11**, 1928 (2020).
37. Derryberry, E. P., Derryberry, G. E., Maley, J. M. & Brumfield, R. T. HZAR: hybrid zone analysis using an R software package. *Mol. Ecol. Resour.* **14**, 652–663 (2014).
38. Hendry, A. P. Selection against migrants contributes to the rapid evolution of ecologically dependent reproductive isolation. *Evol. Ecol. Res.* **6**, 1219–1236 (2004).
39. Nosil, P., Vines, T. H. & Funk, D. J. Perspective: Reproductive isolation caused by natural selection against immigrants from divergent habitats. *Evolution* **59**, 705–719 (2005).
40. Moser, D., Frey, A. & Berner, D. Fitness differences between parapatric lake and stream stickleback revealed by a field transplant. *J. Evol. Biol.* **29**, 711–719 (2016).
41. Berner, D., Roesti, M., Hendry, A. P. & Salzburger, W. Constraints on speciation suggested by comparing lake-stream stickleback divergence across two continents. *Mol. Ecol.* **19**, 4963–4978 (2010).
42. Raeymaekers, J. A. et al. Testing for mating isolation between ecotypes: laboratory experiments with lake, stream and hybrid stickleback. *J. Evol. Biol.* **23**, 2694–2708 (2010).
43. Berner, D. et al. Sexual isolation promotes divergence between parapatric lake and stream stickleback. *J. Evol. Biol.* **30**, 401–411 (2017).
44. Eizaguirre, C., Lenz, T. L., Kalbe, M. & Milinski, M. Divergent selection on locally adapted major histocompatibility complex immune genes experimentally proven in the field. *Ecol. Lett.* **15**, 723–731 (2012).
45. Barton, N. H. & De Cara, M. A. R. The evolution of strong reproductive isolation. *Evolution* **63**, 1171–1190 (2009).
46. Bierne, N., Welch, J., Loire, E., Bonhomme, F. & David, P. The coupling hypopthesis: why genome scans may fail to map local adaptation genes. *Mol. Ecol.* **20**, 2044–2072 (2011).
47. Barton, N. H. Multilocus clines. *Evolution* **37**, 454–471 (1983).
48. Barton, N. H. & Bengtsson, B. O. The barrier to genetic exchange between hybridizing populations. *Heredity* **57**, 357–376 (1986).
49. Flaxman, S. M., Wacholder, A. C., Feder, J. L. & Nosil, P. Theoritical models of the influence of genomic architecture on the dynamics of speciation. *Mol. Ecol.* **23**, 4074–4088 (2014).
50. Fisheries and Ocean Canada. *Recovery strategy for the Misty Lake Sticklebacks (Gasterosteus aculeatus) in Canada* (Species at risk Act Recovery Strategy Series, Canada, 2018).
51. Haenel, Q., Laurentino, T. G., Roesti, M. & Berner, D. Meta-analysis of chromosome-scale crossover rate variation in eukaryotes and its significance to evolutionary genomics. *Mol. Ecol.* **27**, 2477–2497 (2018).
52. Roesti, M., Moser, D. & Berner, D. Recombination in the threespine stickleback genome—patterns and consequences. *Mol. Ecol.* **22**, 3014–3027 (2013).
53. Roesti, M., Hendry, A. P., Salzburger, W. & Berner, D. Genome divergence during evolutionary diversification as revealed in replicate lake–stream stickleback population pairs. *Mol. Ecol.* **21**, 2852–2862 (2012).
54. Berner, D. & Roesti, M. Genomics of adaptive divergence with chromosome-scale heterogeneity in crossover rate. *Mol. Ecol.* **26**, 6351–6369 (2017).
55. Hohenlohe, P. A. et al. Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLoS Genet.* **6**, e1000862 (2010).
56. Deagle, B. E., Jones, F. C., Absher, D. M., Kingsley, D. M. & Reimchen, T. E. Phylogeography and adaptation genetics of stickleback from the Haida Gwaii archipelago revealed using genome-wide single nucleotide polymorphism genotyping. *Mol. Ecol.* **22**, 1917–1932 (2013).
57. Roesti, M., Gavrilets, S., Hendry, A. P., Salzburger, W. & Berner, D. The genomic signature of parallel adaptation from shared genetic variation. *Mol. Ecol.* **23**, 3944–3956 (2014).
58. Nelson, T. C. & Cresko, W. A. Ancient genomic variation underlies repeated ecological adaptation in young stickleback populations. *Evol. Lett.* **2**, 9–21 (2018).
59. Haenel, Q., Roesti, M., Moser, D., MacColl, A. D. C. & Berner, D. Predictable genome-wide sorting of standing genetic variation during parallel adaptation to basic versus acidic environments in stickleback fish. *Evol. Lett.* **3**, 28–42 (2019).
60. Rieseberg, L. H. Chromosomal rearrangements and speciation. *Trends Ecol. Evol.* **16**, 351–358 (2001).
61. Kirkpatrick, M. & Barton, N. Chromosome inversions, local adaptation and speciation. *Genetics* **173**, 419–434 (2006).
62. Yeaman, S. Genomic rearrangements and the evolution of clusters of locally adaptive loci. *Proc. Natl Acad. Sci. USA* **110**, 1743–1751 (2013).
63. Oke, K. B. et al. Does plasticity enhance or dampen phenotypic parallelism? A test with three lake-stream stickleback pairs. *J. Evol. Biol.* **29**, 126–143 (2016).
64. Adams, D. C. & Otárola-Castillo, E. geomorph: an R package for the collection and analysis of geometric morphometric shape data. *Methods Ecol. Evol.* **4**, 363–399 (2013).
65. R Core Team. *A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2019).
66. Rohlf, F. J. & Slice, D. Extensions of the procrustes method for the optimal superimposition of landmarks. *Syst. Zool.* **39**, 40–59 (1990).
67. Moore, J.-S. & Hendry, A. P. Both selection and gene flow are necessary to explain adaptive divergence: evidence from clinal variation in stream stickleback. *Evol. Ecol. Res.* **7**, 871–886 (2005).
68. Gautier, M. et al. Estimation of population allele frequencies from next-generation sequencing data: pool-versus individual-based genotyping. *Mol. Ecol.* **22**, 3766–3779 (2013).
69. Glazer, A. M., Killingbeck, E. E., Mitros, T., Rokhsar, D. S. & Miller, C. T. Genome assembly improvement and mapping convergently evolved skeletal traits in sticklebacks with genotyping-by-sequencing. *G3-Genes Genom. Genet.* **5**, 1463–1472 (2015).
70. Morgan, M., Pagès, H. & Obenchain, V. H. N. *Rsamtools: binary alignment (BAM), FASTA, variant call (BCF), and tabix file import. R package version 2.2.3.* http://bioconductor.org/packages/Rsamtools (2017).
71. Roesti, M., Salzburger, W. & Berner, D. Uninformative polymorphisms bias genome scans for signatures of selection. *BMC Evol. Biol.* **12**, 94 (2012).

Acknowledgements

This project was supported financially by the Swiss National Science Foundation (grant 31003A_165826 to DB) and the Freiwillige Akademische Gesellschaft Basel (QH), and field sampling additionally by Fisheries and Oceans Canada, the British Columbia Ministry of the Environment, and the McGill University Biology Department. We thank Fiona Beaty, Brody Forst, Bailey Feddersen, Tristan Kosciuch, Minxin Lu, Erica MacClaren, Emily McIntosh, Alexanne Oke, Sarah Sanderson, and Mac Willing for aiding field sampling; Western Forest Products for providing logistical and safety support and access to field sites; Walter Salzburger for sharing web lab infrastructure; Brigitte Aeschbach and Nicolas Boileau for facilitating lab work; Christian Beisel, Ina Nissen and Elodie Burcklen for Illumina sequencing at the Quantitative Genomics Facility, D-BSSE, ETH Zürich; the developers of Novocraft for sharing their sequence aligner; Katja Räsänen for providing pictures of Misty stickleback; Laurent Guerard and Nicolás Lichilín Ortiz for help with scripting. Computation was performed at the sciCORE (https://scicore.unibas.ch) scientific computing center of the University of Basel.

Author contributions

D.B. initiated and supervised the study; D.B. and Q.H. designed the experiment; D.B., Q.H., and A.P.H acquired funding; K.B.O. and A.P.H. performed field sampling and measurements; K.B.O. generated and analyzed phenotypic data; Q.H. and T.G.L. performed wet lab work; D.B. and Q.H. implemented analytical tools; Q.H. and D.B. analyzed genomic data and interpreted results; Q.H. and D.B. wrote the manuscript, with input from all co-authors.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-021-25039-y.

Correspondence and requests for materials should be addressed to Q.H. or D.B.

Peer review information *Nature Communications* thanks Mark Ravinet, Zachariah Gompert and the other anonymous reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission information is available at http://www.nature.com/reprints

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

Supplementary Information for:

Clinal genomic analysis reveals strong reproductive isolation across a steep habitat transition in stickleback fish

Quiterie Haenel[1]*, Krista B. Oke[2,3], Telma G. Laurentino[1], Andrew P. Hendry[3] and Daniel Berner[1]*

[1] Department of Environmental Sciences, Zoology, University of Basel, Basel, Switzerland

[2] College of Fisheries and Ocean Sciences, University of Alaska Fairbanks, Juneau, Alaska

[3] Redpath Museum and Department of Biology, McGill University, Montreal, Quebec, Canada

*corresponding authors: quiterie.haenel@unibas.ch, daniel.berner@unibas.ch

Contents

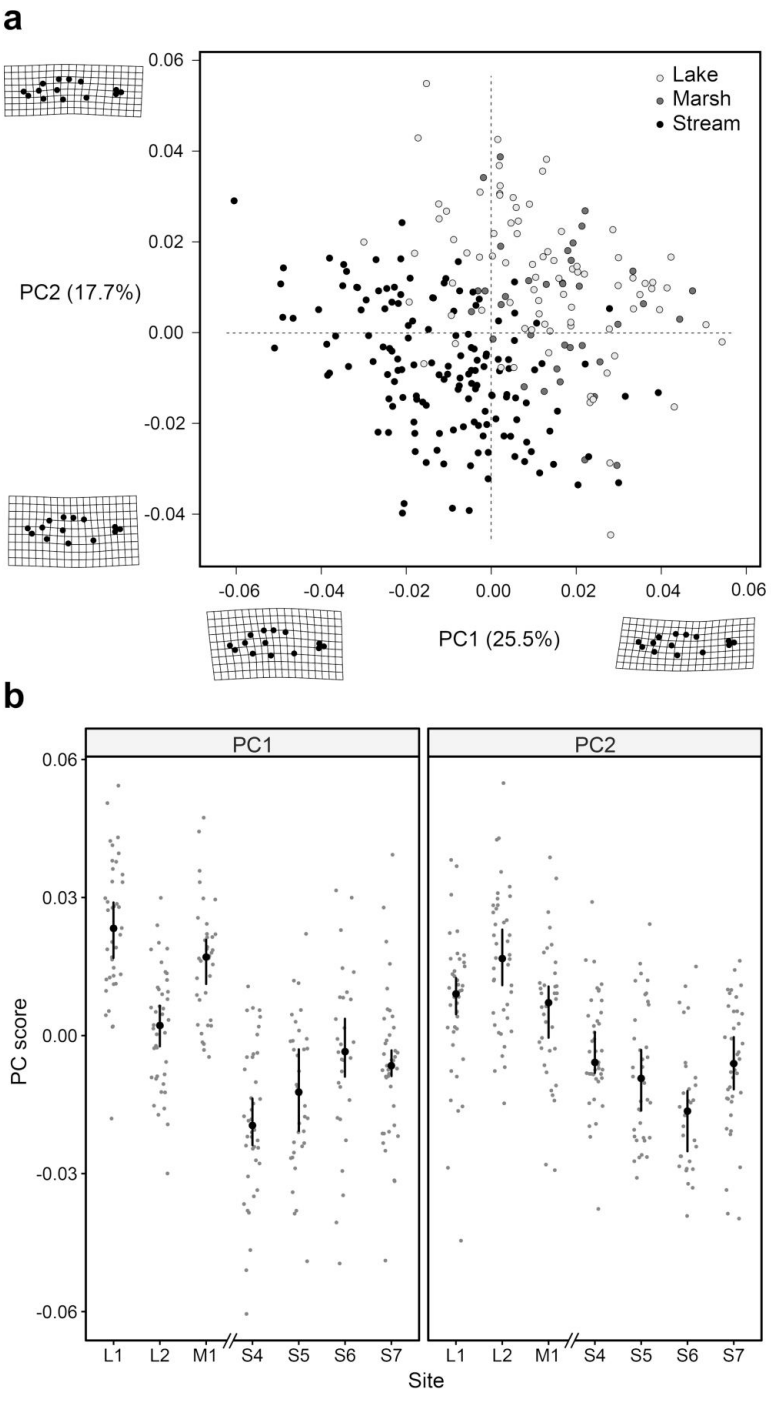
Supplementary Figures

Supplementary Tables

Supplementary References

Supplementary Figures

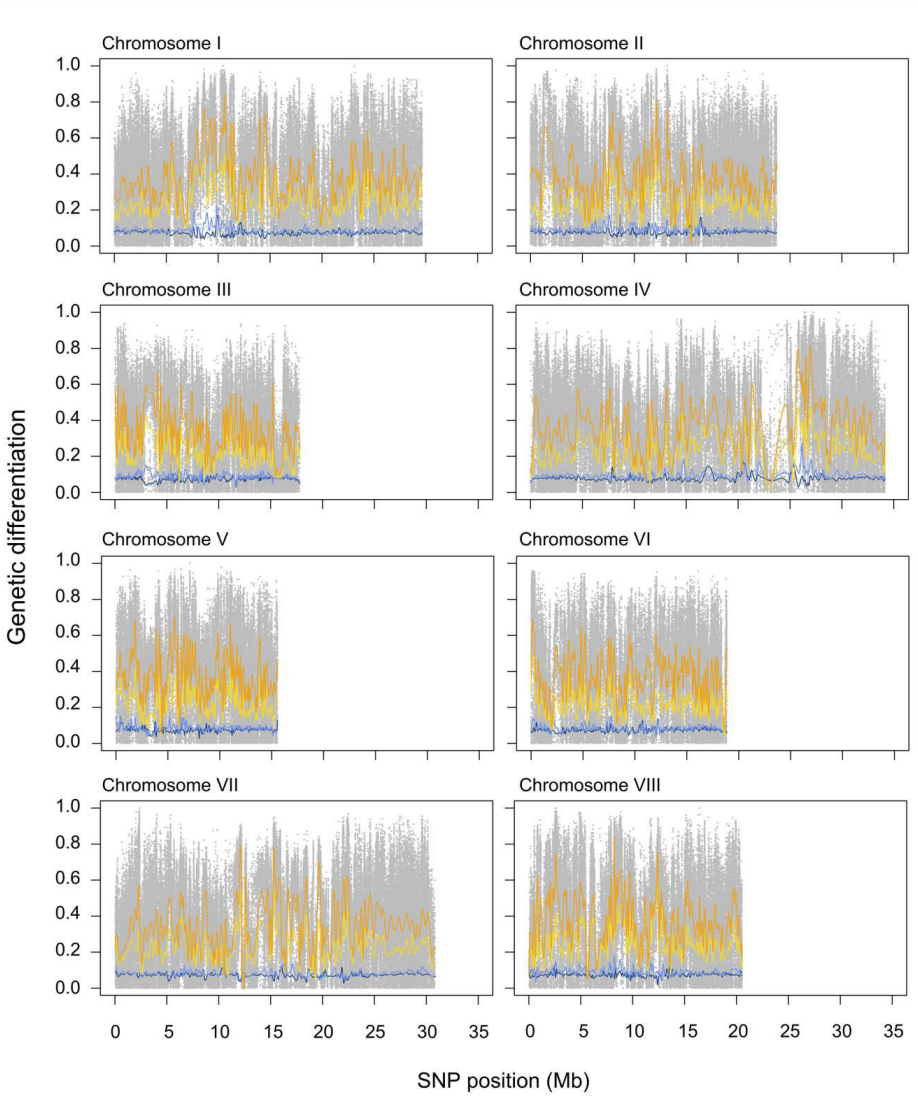
Supplementary Fig. 1



4

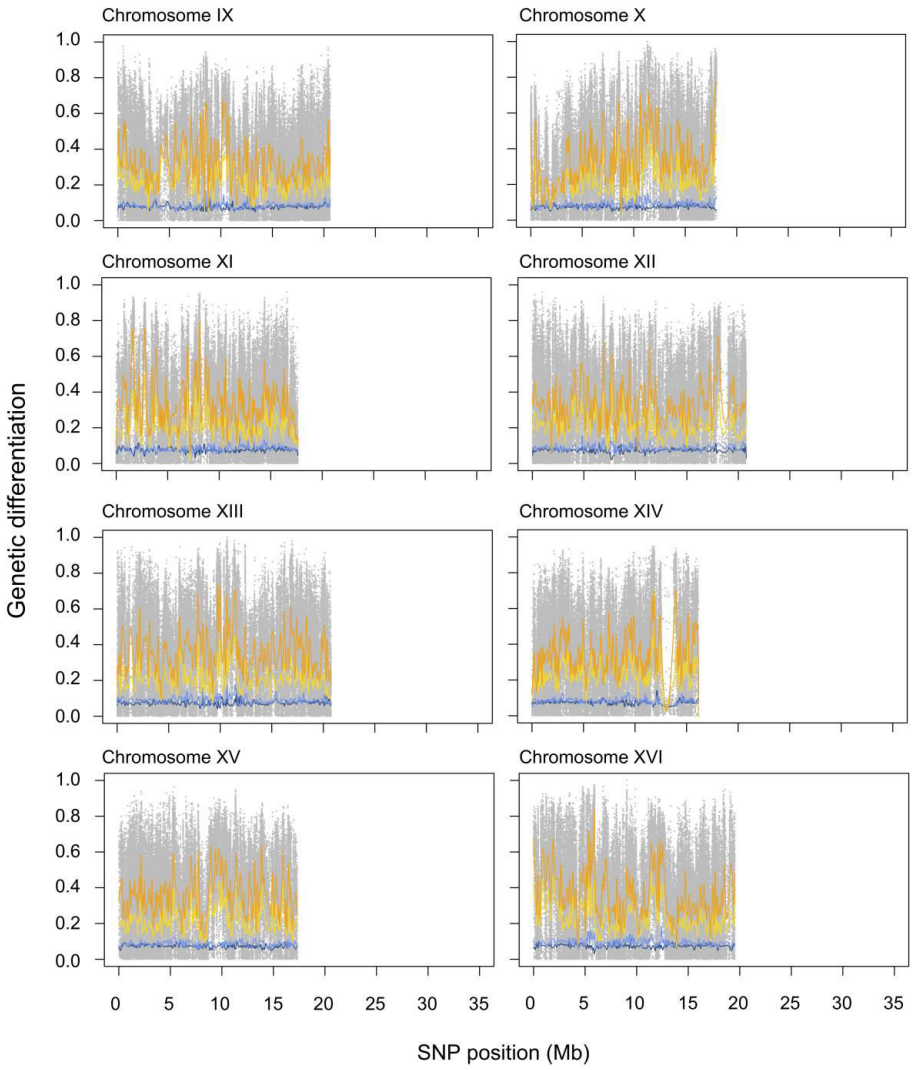
Supplementary Fig. 1 Geometric morphometric analysis of lake, marsh and stream stickleback. a Each point represents an individual fish along the first two principal components (PC1, PC2) obtained by landmark-based shape analysis (methodological details given in refs. [1,2]). Individuals are color coded according to their habitat (lake, marsh, inlet stream). The deformation grids visualize the body shape associated with the lowest and highest observed score along each PC. **b** Individual PC scores shown separately for each study site (n = 40 individuals per site). Black dots and vertical lines represent site medians with their bootstrap 95% compatibility interval. Note that both PCs capture variation in body depth, that stream fish tend to exhibit deeper bodies than lake fish, and that marsh fish resemble lake fish.

Supplementary Fig. 2

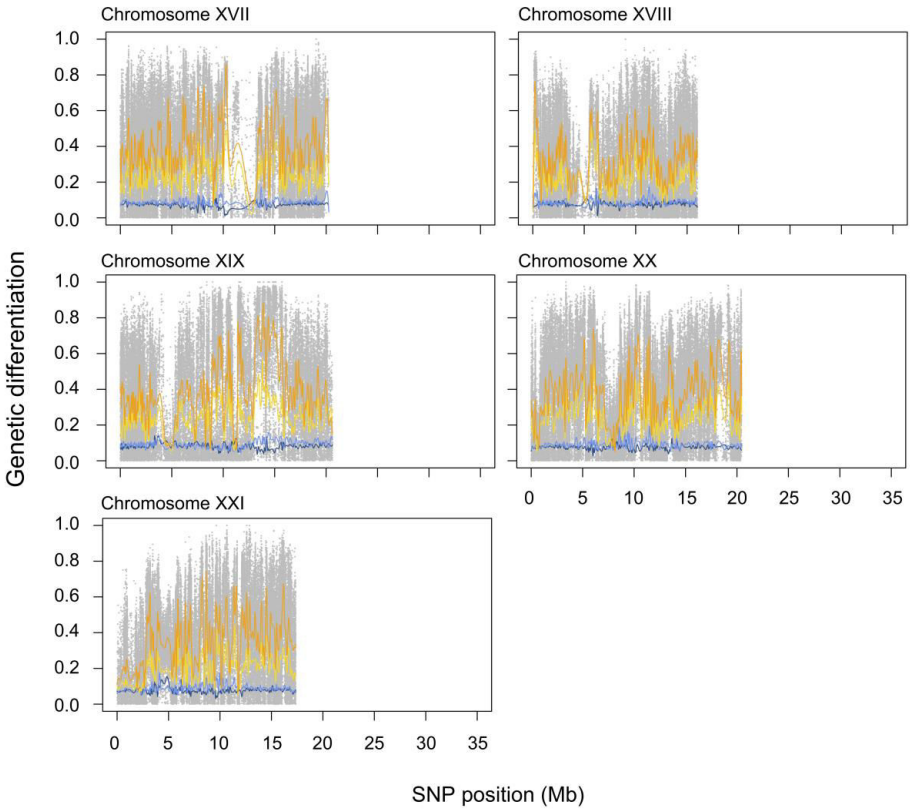


6

Supplementary Fig. 2

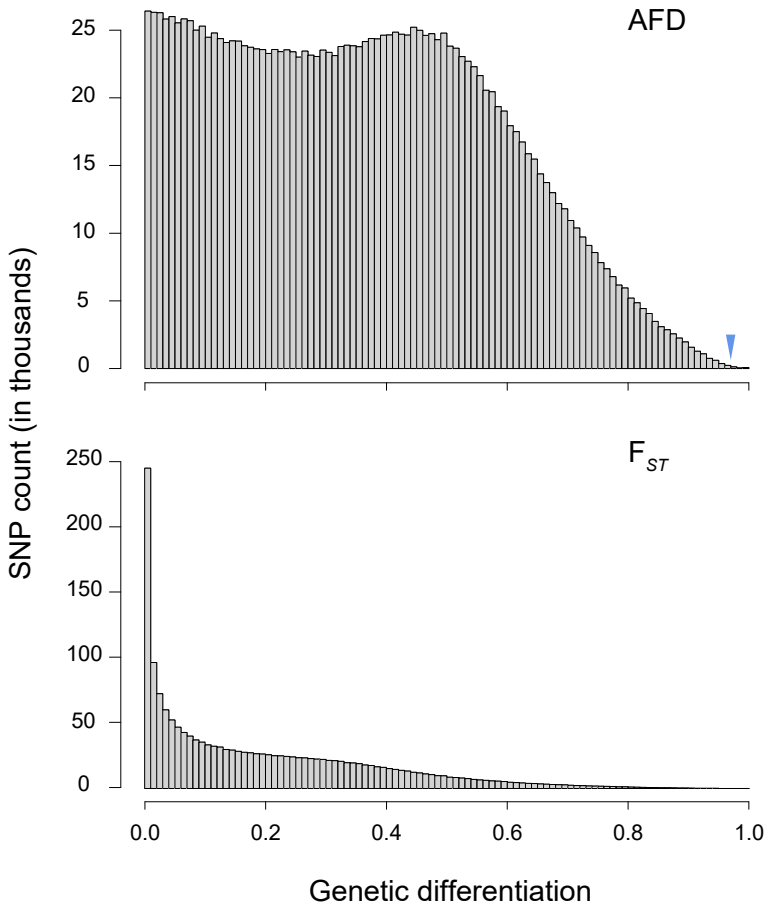


Supplementary Fig. 2



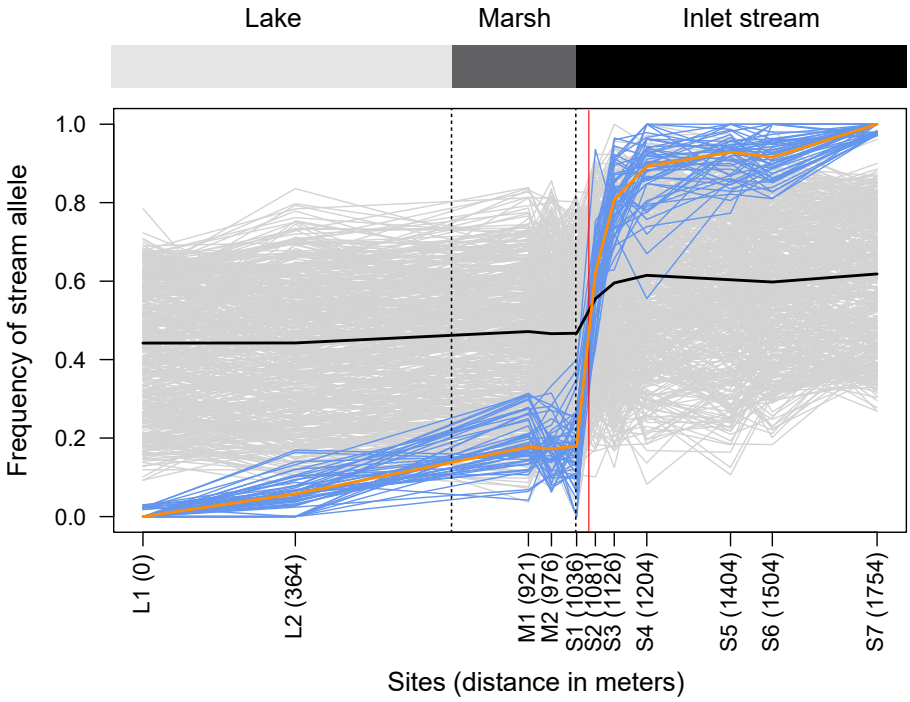
Supplementary Fig. 2 Genetic differentiation between Misty Lake and inlet stream stickleback across all chromosomes. Differentiation is expressed by the absolute allele frequency difference AFD. The presentation format follows that of Fig. 3.

Supplementary Fig. 3



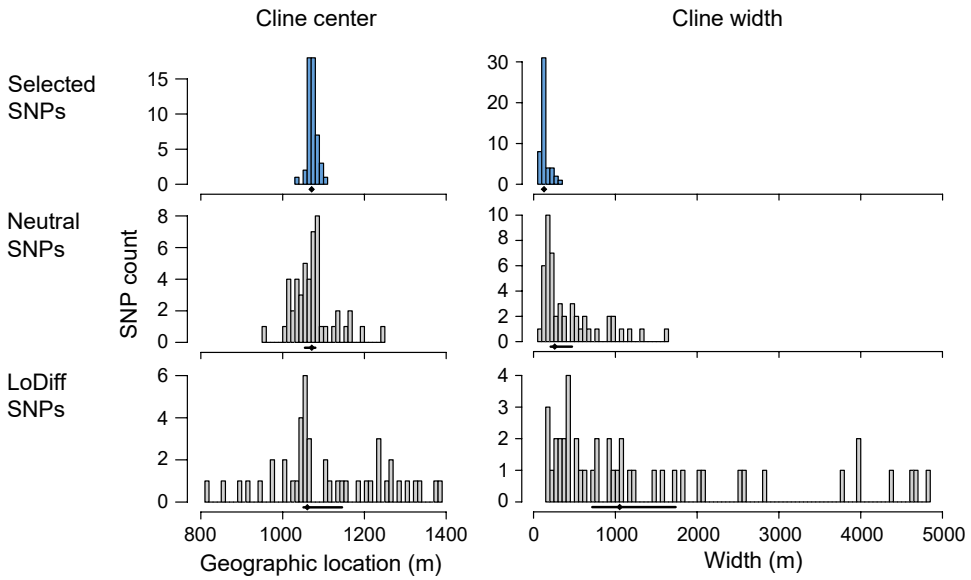
Supplementary Fig. 3 Distribution of differentiation between the sites L1 and S7 across all genome-wide SNPs. In the top panel, differentiation at 1,708,118 SNPs (including the sex chromosome) passing quality filtering thresholds for this specific sample comparison is expressed by the absolute allele frequency difference AFD. To facilitate comparison with previous work, the lower panel shows the analogous distribution based on an F_{ST} estimator (G_{ST}[3]). The genome-wide median differentiation is 0.352 (AFD) and 0.139 (F_{ST}). The blue triangle in the upper panel indicates the threshold AFD value of 0.97 that was applied to identify the panel of selected SNPs.

Supplementary Fig. 4



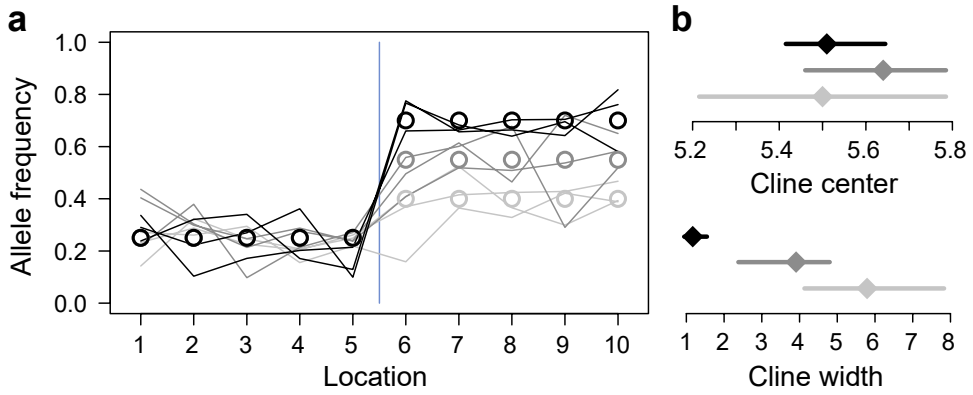
Supplementary Fig. 4 Genomic differentiation along the lake-stream transition at the loDiff SNPs. This graphic is analogous to Fig. 2b, except that the allele frequencies from the neutral SNP category are here replaced by the corresponding data from loDiff SNPs (gray and black lines). The loDiff SNPs are 500 markers chosen at random among all SNPs deviating by no more than 0.1% from half the genome-wide median AFD (c. 0.17) in the L1-S7 sample comparison. These SNPs are even less likely than the neutral SNPs to be influenced by divergent selection on physically close genome regions, hence capture genome-wide differentiation by drift between Misty Lake and inlet stream stickleback.

Supplementary Fig. 5



Supplementary Fig. 5 Raw distribution of cline center location and cline width estimates. The histograms show cline center and width estimates obtained by genetic cline modeling for each of the 50 individual SNPs from the three marker categories (selected SNPs in blue). The underlying data points are the medians across ten replicate model fitting runs for each SNP. Below each histogram, the median and the associated 95% bootstrap compatibility interval across the SNPs are visualized (values identical to those in Fig. 4). A single SNP from the neutral SNP panel displayed extremely spatially unstable allele frequencies compared to all other SNPs and was therefore considered a technical outlier and excluded from the histogram (center estimate 1301 m, width estimate 2701 m), although this SNP was included for the summary statistics (having a trivial influence). Also, to maintain a reasonably high visual resolution along the X-axis, five loDiff SNPs exhibiting extreme values for cline center (-458 m, 368 m, 431 m, 1516 m, 1655 m) and cline width (5229 m, 5321 m, 5471 m, 7919 m, 20766 m) were omitted from the histograms, although these SNPs were not considered technical outliers and were included for the calculation of summary statistics.

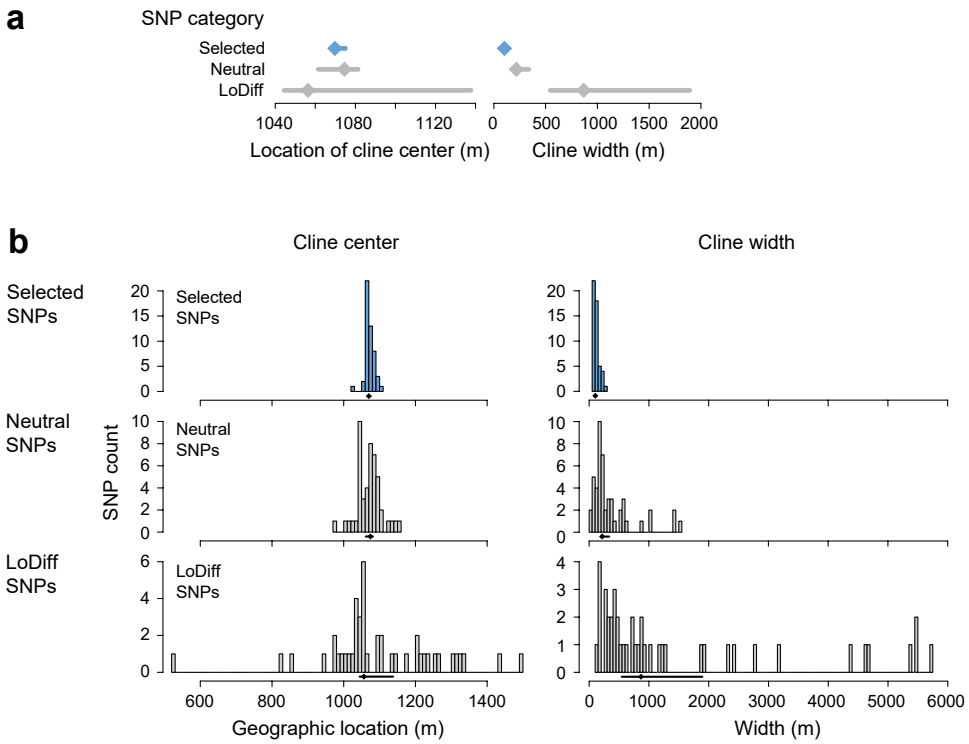
Supplementary Fig. 6



Supplementary Fig. 6 Simulation study to assess the influence of the magnitude of differentiation between two contiguous populations on cline parameter estimation. a We modeled two populations, each represented by five evenly spaced geographic locations. The boundary between the populations thus occurred between the locations 5 and 6 (vertical blue line) and coincided with a shift in expected allele frequencies. The magnitude of this shift (AFD) differed among simulations from 0.15 (light gray) to 0.30 (gray) and 0.45 (black; stronger differentiation would have produced allele frequencies exceeding the 0-1 range and was therefore avoided). The population on the left of the allele frequency breakpoint always had an expected frequency of 0.25, while the expected frequency of the population on the right of the breakpoint was obtained by adding the corresponding AFD (the expected frequencies at each location are shown as circles; on the left of the breakpoint, these are perfectly overlapping). To introduce stochasticity, we added a random draw from the normal distribution with a mean of zero and a standard deviation of 0.08 to the expected allele frequency at each site (this standard deviation produced variation in allele frequencies among locations qualitatively similar to the variation observed empirically). The expected magnitude of random noise and the true cline position and width were thus tightly controlled and invariant among the different simulation types, which differed only in AFD between the populations. The final allele frequency data obtained in this way (three randomly chosen examples are illustrated by lines for each AFD category) were then subject to cline parameter

estimation with HZAR[4], analogously to our empirical analysis. We here used the locations as distance values and consistently assumed a sample size of 100 underlying allele frequency estimation at any given location. The estimates for cline center and width from a single MCMC run per simulation replicate were then saved, with 30 such replicates run per AFD category. Data from the replicate simulations were used to compute median cline center and width and their 95% bootstrap compatibility intervals (CIs) based on 10,000 resamples. **b** This simulation experiment produced insights relevant to the interpretation of our empirical analysis: first, we found that across all three modeled levels of AFD between the two populations, cline center location was estimated accurately (top panel); the median estimate approximated the expected value of 5.5 closely, or the latter was at least well within the CI. However, estimation precision for cline center decreased (wider CIs) with decreasing AFD between the populations. The second observation was that median cline width and its spread increased with decreasing AFD (bottom panel). All these observations mirror patterns also emerging from our empirical analysis (Fig. 4). This leads us to conclude that the greater cline widths observed empirically for the neutral and especially the loDiff SNPs relative to the selected SNPs offer no evidence of geographically more extensive gene flow in the latter SNP categories. Even if reproductive isolation is complete, genome regions exhibiting greater population differentiation by chance (stronger drift) or due to divergent selection will produce lower cline width estimates. We note that this influence of the magnitude of AFD on cline parameters was observed even when fitting models to the precise population-specific allele frequencies from all locations (i.e., no random noise), when running models without tails, and when increasing the number of geographic locations on either side of the population boundary to ten (details not presented).

Supplementary Fig. 7

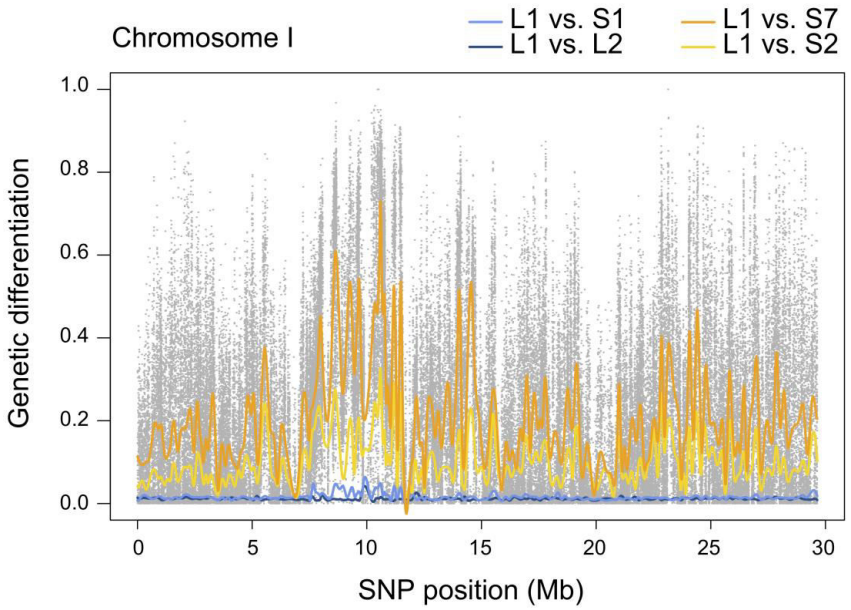


Supplementary Fig. 7 Genetic cline modeling with the two terminal sites of the geographic gradient excluded. To examine potential SNP ascertainment bias in cline model fitting, all modeling was here repeated with allele frequency data from the sampling sites L1 and S7 excluded, resulting in nine total sampling sites. The rationale was that AFD observed in the comparison of these two sites was used as the basis for delimiting the SNPs for the selected, neutral, and loDiff marker panels. The allele frequencies at L1 and S7 were thus not fully independent, contrary to all other sampling sites. Cline modeling with these reduced data sets was otherwise carried out as described for the analysis with the full 11 sites, and the same graphing conventions were followed (see Fig. 4 and Supplementary Fig. 5). Both the summary statistics **a** and the underlying distributions **b** of cline center and width estimates across SNPs for the three SNP categories are very similar to the corresponding

results with the sites L1 and S7 included (Fig. 4 and Supplementary Fig. 5) and support the same conclusions. In the histograms for the neutral SNPs, a single technical outlier marker was again excluded (estimated cline center 7099 m; cline width 140705 m), and for ease of presentation, a few loDiff SNPs showing extreme estimates for cline center (-2928 m, 254 m, 1711 m, 1946 m, 6086 m) and width (6226 m, 6341 m, 6559 m, 19400 m, 29821 m, 57318 m) are not visualized. However, all these SNPs were included for the calculation of the summary statistics in **a**.

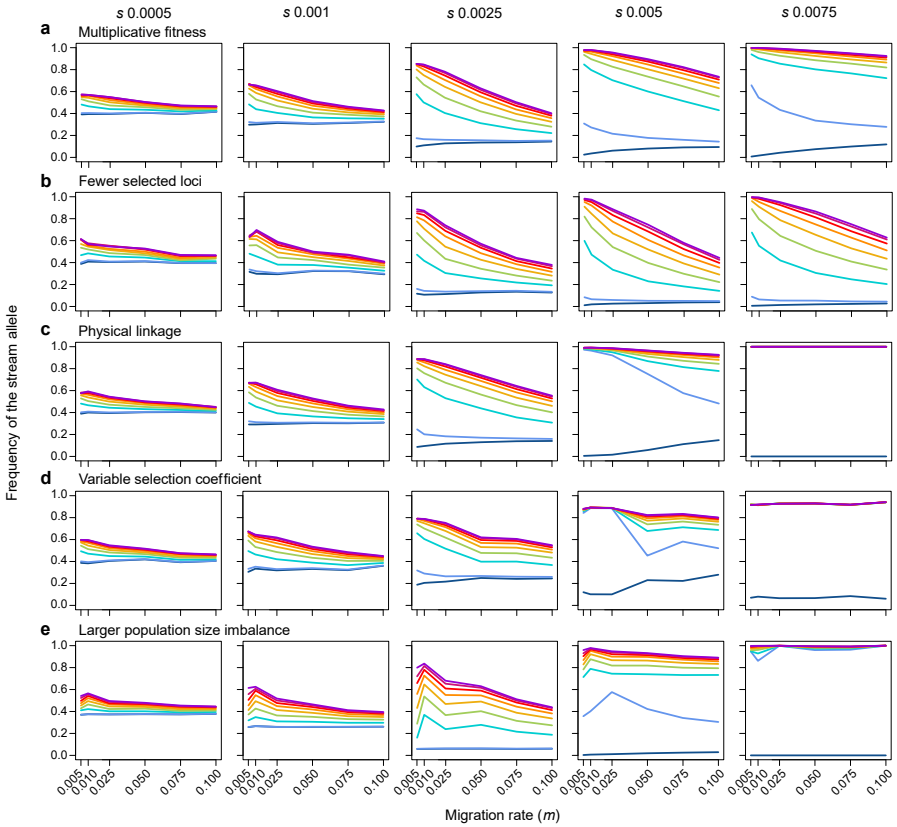
15

Supplementary Fig. 8



Supplementary Fig. 8 Pairwise differentiation along a chromosome, expressed by F_{ST}. The figure follows the same conventions as Fig. 3, except that differentiation at each SNP in each sampling site comparison was quantified by F_{ST} (the G_{ST} estimator of ref. [3]). Note that due to the low sensitivity of F_{ST} compared to AFD when population differentiation is weak or modest[5], the relatively strong differentiation peaks near 10 Mb in the L1-S1 site comparison are less obvious when differentiation is based on F_{ST} as opposed to AFD (compare to Fig. 3).

Supplementary Fig. 9

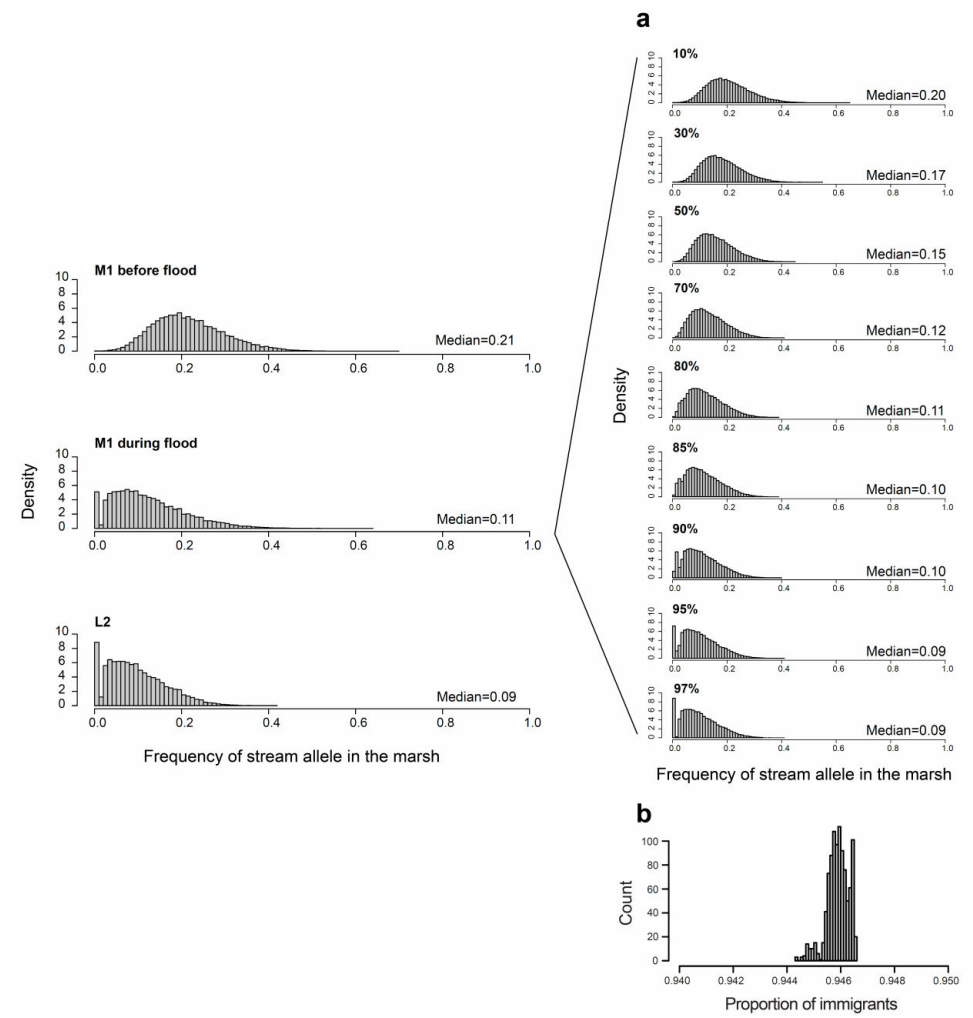


Supplementary Fig. 9 Robustness checks of the simulations of divergence with gene flow across a habitat transition. The presentation format follows Fig. 6, but the simulations were performed with the standard model modified in the following ways: **a** The loci contribute to fitness multiplicatively, as opposed to additively. **b** Only ten loci are under selection, as opposed to 100. **c** All loci are physically linked on a single chromosome exhibiting crossover, as opposed to free segregation. **d** The selection coefficients are not identical among loci, but are drawn from an exponential distribution (rate = 1/*s*). Note that the latter causes greater stochasticity in the simulation outcome, as particularly evident with s = 0.005. **e** The lake population size is ten times the stream population size.

Supplementary Fig. 10

Supplementary Fig. 10 Characterization of the marsh habitat in the Misty system. The marsh represents the transition from the inlet stream (flowing through woodland) to the lake. This habitat is dominated by short (< 1.5 m), dense vegetation intersected by relatively narrow, deep channels. During most of the summer, one can walk on mats of this vegetation, the roots of which reach approximately 30 cm below the water surface. During the flood, these mats became entirely submerged, likely allowing fish to swim through vegetation that days before would have been above water. Supporting this view, stickleback catch rates at the marsh site increased dramatically during the flood (Krista B. Oke, unpublished data) (Photo credits: Krista B. Oke).

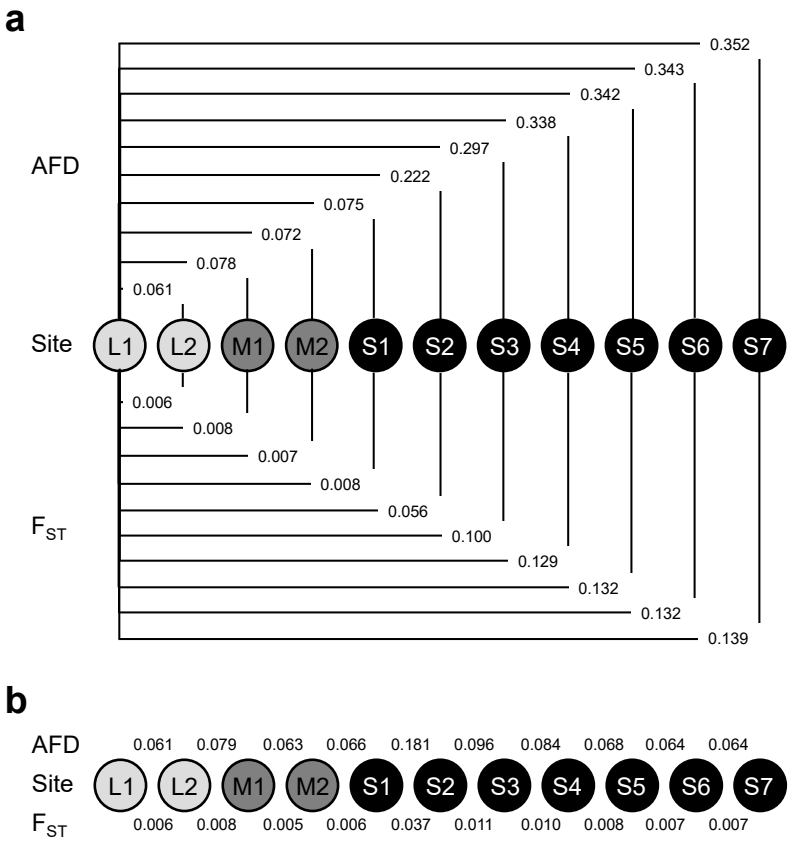
Supplementary Fig. 11



Supplementary Fig. 11 Exploring the approximate proportion of migrants from the lake into the marsh during the flood. This analysis focused on the same 49,677 SNPs highly differentiated between the lake and stream population also underlying Fig. 7. We here performed weighted averaging of the frequency of the stream allele at each SNP between the marsh sample (M1) before the flood (top left histogram) and the lake sample closest to this marsh site (L2; bottom left). The relative weight of each sample was varied to mimic different levels of dispersal of

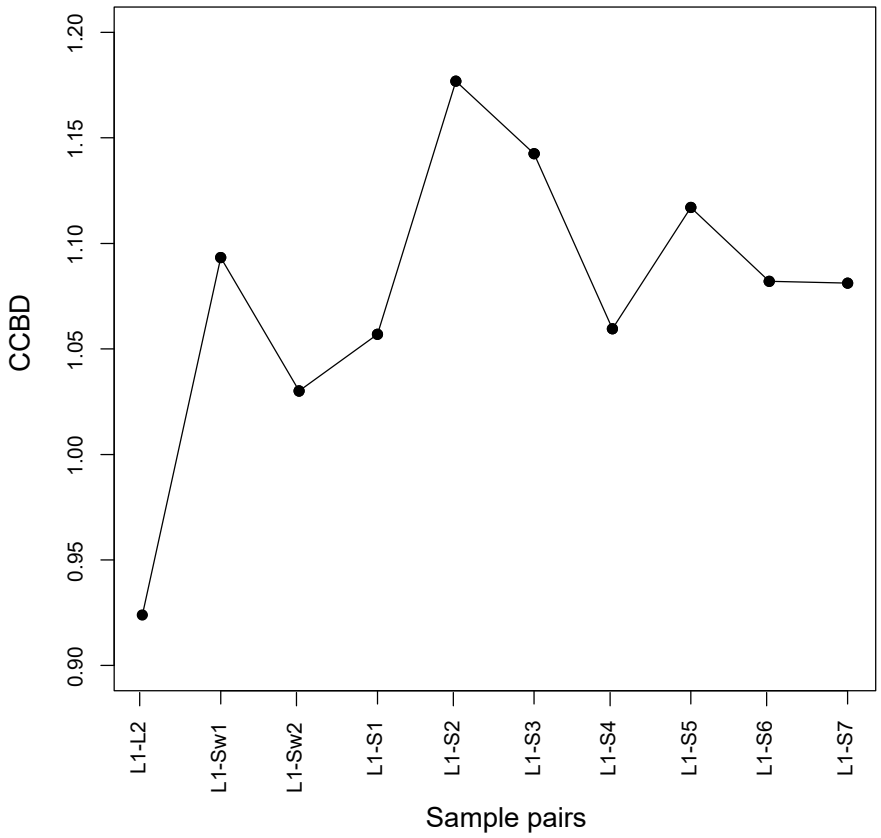
lake fish into the marsh during the flood. We then asked what relative proportion of immigrant lake fish at the marsh site is required to yield a stream allele frequency distribution qualitatively resembling the distribution observed empirically during the flood (middle left). The evaluation of the resulting distributions was first performed visually, paying particular attention to the proportion of SNPs exhibiting a stream allele frequency very near zero. The allele frequency distributions resulting from nine exemplary immigrant proportions (indicated on top of each panel) are shown in **a**. This exploration suggested that the proportion of lake immigrants present at the marsh site during the flood was very high, likely around 90 - 95%. In addition, this proportion was estimated by approximate Bayesian computation (ABC). We here characterized the stream allele frequency distribution by using its 0.1, 0.5 and 0.9 quantiles as summary statistics. These statistics were calculated for 10,000 iterations, each using a relative proportion of immigrant lake fish drawn at random from the uniform distribution bounded between 0 and 1. The estimation of the proportion parameter was then performed with the *abc* R package[6], using a tolerance of 0.1 and the neural networks-based method for constructing the posterior distribution. The weighted median of the posterior distribution shown in **b** was 94.6% (2.5 and 97.5 percentiles: 0.9455, 0.9465), in good agreement with our visual estimation. Qualitatively similar results were obtained when using the simple rejection algorithm for generating the posterior distribution, and across the full range of tolerance values explored (0.2 - 0.01); the median of the posterior distribution was always between 0.80 and 0.95.

Supplementary Fig. 12



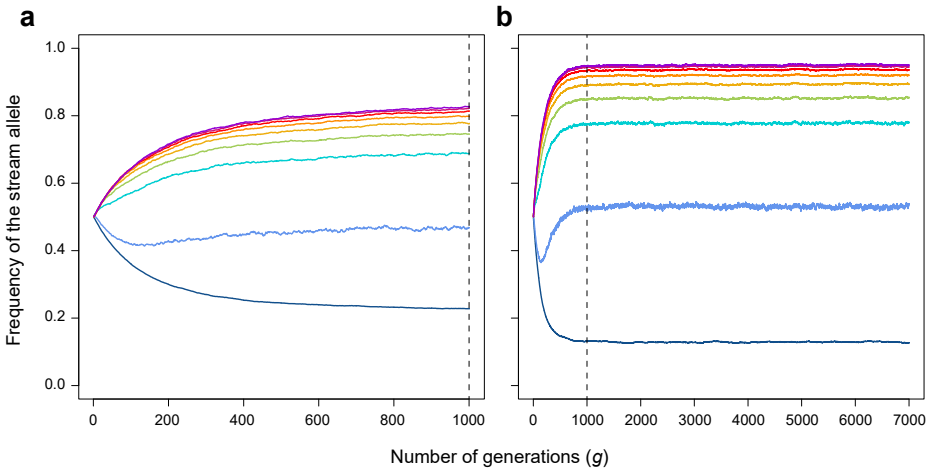
Supplementary Fig. 12 Genetic differentiation between the study sites. Median genetic differentiation, expressed by the absolute allele frequency difference AFD and F_{ST} (the estimator G_{ST} of ref. [3]), across all genome-wide SNPs (including the sex chromosome). Differentiation was calculated for all pairwise comparisons between L1 and each of the other sample sites **a**, and for all pairwise combinations of neighboring sample sites along the geographic gradient **b**.

Supplementary Fig. 13



Supplementary Fig. 13 Alternative analysis of chromosome center-biased differentiation (CCBD). CCBD is here calculated as described in the Methods, but the samples used for pairwise comparison were not from adjacent sites as in Fig. 2c, but involved all combinations of the samples L2 to S7 with the sample L1. Consistent with Fig. 2c, the magnitude of CCBD is greatest for the L1-S2 sample pair, indicating selection-gene flow antagonism in the lowest reach of the inlet stream.

Supplementary Fig. 14



Supplementary Fig. 14 Determining an appropriate number of generations for the individual-based simulations. Simulations were run with our standard stepping stone model over **a** 1000 and **b** 7000 generations for an exemplary migration rate and selection coefficient combination (*m* 0.05 and *s* 0.005). Shown is the frequency of the stream allele over time averaged over 20 simulation replications. The dotted line indicates generation 1000 in both graphs. This exploration indicates that running our simulation model over 1000 generations allows the system to approach migration-selection balance.

Supplementary Tables

Supplementary Table 1 Characterization of the study sites in the Misty Lake watershed. Habitat type, GPS coordinates (in decimal degrees), and the number of individuals for the genomic and morphometric analyses are given for each site. For the genomic sample sizes, the values in parentheses indicate median read depth across all genome-wide positions. The site M1 was sampled at three different time points.

Site	Habitat	Latitude	Longitude	N genomics	N morphometrics
L1	Lake	50.60507824	-127.2685989	62 (103)	40
L2	Lake	50.604347	-127.262569	56 (80)	42
M1	Marsh	50.60516595	-127.2579478	50 (133)	36
M1 During the flood				56 (79)	-
M1 One year later				56 (114)	-
M2	Marsh	50.605087	-127.257812	56 (106)	-
S1	Stream	50.604618	-127.257198	56 (74)	-
S2	Stream	50.604414	-127.256683	56 (93)	-
S3	Stream	50.604375	-127.256141	56 (51)	-
S4	Stream	50.603808	-127.255397	40 (103)	41
S5	Stream	50.603056	-127.252444	52 (120)	37
S6	Stream	50.60223555	-127.2507798	56 (112)	33
S7	Stream	50.60060871	-127.2476535	50 (107)	41

Supplementary References

1. Kaeuffer, R., Peichel, C. L., Bolnick, D. I. & Hendry, A. P. Parallel and nonparallel aspects of ecological, phenotypic, and genetic divergence across replicate population pairs of lake and stream stickleback. *Evolution* **66**, 402-418 (2012).

2. Oke, K. B., Bukhari, M., Kaeuffer, R., Rolshausen, G., Rasanen, K., Bolnick, D. I. & Hendry, A. P. Does plasticity enhance or dampen phenotypic parallelism? A test with three lake-stream stickleback pairs. *J. Evol. Biol.* **29**, 126-143 (2016).

3. Nei, M. Analysis of gene diversity in subdivided populations. *PNAS USA* **70**, 3321-3323 (1973).

4. Derryberry, E. P., Derryberry, G. E., Maley, J. M. & Brumfield, R. T. HZAR: hybrid zone analysis using an R software package. *Mol. Ecol. Res.* **14**, 652-63 (2014).

5. Berner, D. Allele frequency difference *AFD* – an intuitive alternative to F_{ST} for quantifying genetic population differentiation. *Genes* **10**, 308 (2019).

6. Csilléry, K., François, O. & Blum, M. G. B. abc: an R package for approximate Bayesian computation (ABC). *Methods Ecol. Evol.* **3**, 475–79 (2012).

Chapter 4

Meta-analysis of chromosome-scale crossover rate variation in eukaryotes and its significance to evolutionary genomics

Haenel, Laurentino et al. 2018, Molecular Ecology

Received: 9 November 2017 | Revised: 23 March 2018 | Accepted: 26 March 2018

DOI: 10.1111/mec.14699

INVITED REVIEWS AND SYNTHESES

WILEY MOLECULAR ECOLOGY

Meta-analysis of chromosome-scale crossover rate variation in eukaryotes and its significance to evolutionary genomics

Quiterie Haenel[1]* | Telma G. Laurentino[1]* | Marius Roesti[2] | Daniel Berner[1]

[1]Zoological Institute, University of Basel, Basel, Switzerland

[2]Department of Zoology, University of British Columbia, Vancouver, BC, Canada

Correspondence
Daniel Berner, Zoological Institute, University of Basel, Basel, Switzerland.
Email: daniel.berner@unibas.ch

Funding information
Janggen-Pöhn Foundation; Schweizerischer Nationalfonds zur Förderung der Wissenschaftlichen Forschung, Grant/Award Number: 31003A_165826

Abstract

Understanding the distribution of crossovers along chromosomes is crucial to evolutionary genomics because the crossover rate determines how strongly a genome region is influenced by natural selection on linked sites. Nevertheless, generalities in the chromosome-scale distribution of crossovers have not been investigated formally. We fill this gap by synthesizing joint information on genetic and physical maps across 62 animal, plant and fungal species. Our quantitative analysis reveals a strong and taxonomically widespread reduction of the crossover rate in the centre of chromosomes relative to their peripheries. We demonstrate that this pattern is poorly explained by the position of the centromere, but find that the magnitude of the relative reduction in the crossover rate in chromosome centres increases with chromosome length. That is, long chromosomes often display a dramatically low crossover rate in their centre, whereas short chromosomes exhibit a relatively homogeneous crossover rate. This observation is compatible with a model in which crossover is initiated from the chromosome tips, an idea with preliminary support from mechanistic investigations of meiotic recombination. Consequently, we show that organisms achieve a higher genome-wide crossover rate by evolving smaller chromosomes. Summarizing theory and providing empirical examples, we finally highlight that taxonomically widespread and systematic heterogeneity in crossover rate along chromosomes generates predictable broad-scale trends in genetic diversity and population differentiation by modifying the impact of natural selection among regions within a genome. We conclude by emphasizing that chromosome-scale heterogeneity in crossover rate should urgently be incorporated into analytical tools in evolutionary genomics, and in the interpretation of resulting patterns.

KEYWORDS
centromere, chromosome length, gene density, linked selection, meiosis, recombination

1 | INTRODUCTION

Meiosis is a specialized cell division widely conserved among sexually reproducing eukaryotes, involving one round of DNA replication followed by two rounds of chromosome division, thus producing haploid cells (gametes, spores) from diploid progenitors. During the first meiotic division, homologous chromosomes pair and undergo recombination. This involves numerous programmed DNA double-strand breaks and the invasion of short single-stranded DNA segments into the homologous chromosome. A small fraction of the DNA breaks are then repaired as crossovers (CO), the reciprocal exchange of DNA segments between the homologous chromosomes (Hunter, 2007; CO is thus only one aspect of recombination, and hence, these two terms are not used interchangeably in this study). CO is an intriguing biological process because of its dual mechanistic and evolutionary implications. On the one hand, the segregation of

*Quiterie Haenel and Telma G. Laurentino share first authorship, alphabetical order.

chromosomes during the first meiotic division requires that homologous chromosomes associate physically to align together on the meiotic spindle in the cell's equator, which is facilitated by CO (or chiasma, the cytological manifestation of CO). CO is thus important for proper chromosome disjunction. Although exceptions exist (e.g., the absence of CO in some dipteran males or lepidopteran females; Gerton & Hawley, 2005; Wolf, 1994), one obligate CO per chromosome pair is generally considered a requirement for accurate chromosome segregation and hence the production of genetically balanced offspring (Hassold & Hunt, 2001; Hunter, 2007; Mather, 1938; Smith & Nicolas, 1998).

On the other hand, CO also has crucial evolutionary consequences: by breaking and interchanging DNA segments from two homologous chromosomes, CO generates novel combinations of alleles. A possible benefit of this genetic reshuffling is that favourable alleles initially occurring on different copies of a given chromosome can be unified into a single chromosome. This chromosome combines the selective benefit of all the alleles it carries, and hence represents a genotype of higher fitness than what would be possible in the absence of CO. CO thus increases genetic variance among individuals, therefore making natural selection in finite populations more efficient (Burt, 2000; Felsenstein, 1974; Fisher, 1930; Hartfield & Otto, 2011; Hill & Robertson, 1966; Kondrashov, 1982; Muller, 1932; Otto & Barton, 1997, 2001)—an effect providing a general explanation for the evolutionary benefit of sexual over asexual reproduction. However, the increase in genetic variation due to CO can also entail a reduction in the mean fitness of a population, for instance when favourable epistatic interactions among loci are broken down (Barton, 1995; Fisher, 1930), or when populations adapted to selectively different habitats hybridize and locally favourable and unfavourable alleles become associated (Barton & Bengtsson, 1986; Berner & Roesti, 2017; Kirkpatrick & Barton, 2006; Ortiz-Barrientos, Reiland, Hey, & Noor, 2002).

The evolutionary consequences of CO depend strongly on the distribution of CO along chromosomes. At a fine scale, the CO rate is often dramatically elevated in localized "hotspots" (Baudat, Imai, & de Massy, 2013; Choi & Henderson, 2015; Lichten & Goldman, 1995). While the distribution of hotspots and their molecular control are under intensive investigation, less attention has been paid to the distribution of CO along chromosomes at a broad scale. In several organisms, it has long been noticed that the CO rate differs greatly among broad chromosome regions (Akhunov et al., 2003; Croft & Jones, 1989; International Human Genome Sequencing Consortium 2001; Nachman & Churchill, 1996; Rahn & Solari, 1986; Rees & Dale, 1974), but so far, no attempt has been made to formally examine the distribution of CO at a large chromosomal scale across taxa.

The objective of this study is to fill this gap by exploiting the recent proliferation of well-characterized CO landscapes in higher eukaryotes (animals, plants, fungi) driven by progress in genome sequencing and marker generation techniques. Using a meta-analytical approach, we document a widespread trend of CO to occur at a relatively elevated rate in the chromosome peripheries. We address the mechanisms potentially causing this pattern and highlight why appreciating this nonrandom distribution of CO across the genome is important to evolutionary population genomics.

2 | METHODS

2.1 | Data acquisition

To initiate our meta-analysis, we conducted a literature search for studies characterizing the distribution of CO across the genome. We considered two types of data sets: first, studies reporting the genetic map position of genetic markers along with their physical base pair position along chromosomes, that is, centimorgan (cM) vs. megabase (Mb) data (>80% of the data sets eventually used). Second, we also considered studies directly reporting CO rates along chromosomes quantified as genetic map distance divided by physical map distance for marker intervals (i.e., cM/Mb vs. Mb data). Our focus was on organisms with an assembled genome (in a few cases, this genome was from a close congeneric species) and with CO rates estimated from crosses or pedigrees. Studies estimating CO rates from linkage disequilibrium in population samples, presenting information from a single chromosome only, or performed with low marker resolution (fewer than ~20 markers per chromosome on average) were ignored. In a single case (Nunes et al., 2017), we considered a marker-dense data set presenting genetic map position against marker order (instead of Mb position in a physical assembly). Visually comparing patterns of cM vs. Mb to cM vs. marker order in another organism in which both data types were available (Dohm et al., 2012, 2014) confirmed that the latter data type also reliably captures broad-scale CO patterns (see also Nachman & Churchill, 1996). All species were assigned to the categories "wild" or "domesticated", the latter subsuming all systems at least potentially having experienced selection by humans (i.e., domesticated, cultivated or classical laboratory model organisms). For species in which suitable CO data were available from multiple independent investigations, we prioritized the study with the most reliable genome assembly and/or the highest marker resolution.

In some studies, the relevant raw data were presented directly in tabulated form. Otherwise, we extracted information from graphics using *webplotdigitizer* (http://arohatgi.info/WebPlotDigitizer). In graphics permitting the identification of individual raw data points, the latter were digitized directly. When marker resolution was relatively sparse, we considered all available data points (ignoring obvious outliers caused by genome misassembly). In high-resolution studies with heavily overlapping data points, we digitized only a subset of points per chromosome sufficient to capture broad-scale CO patterns accurately (a few such data sets digitized independently by multiple researchers confirmed that this subsampling produced highly reproducible CO rate data). In cases where the data were presented as line graphics (e.g., smoothed profiles along chromosomes), and hence, the raw data were not accessible, we superposed a grid of equidistant lines orthogonal to the Mb axis on the plot of each chromosome and digitized the intersections between grid and data lines. This grid was adjusted to span the entire chromosome and included either 26 lines for cM vs. Mb plots, or 25 lines for cM/Mb vs. Mb

plots, eventually yielding CO rate estimates for a minimum of 25 windows along each chromosome across all studies. In studies providing CO information separately for both sexes or for multiple crosses, data were extracted separately for each category and then averaged for analysis (note that in *Drosophila*, CO occurs only in females; we nevertheless considered this species for analysis, although excluding it did not influence any conclusion). To avoid bias by unusual patterns of CO in sex chromosomes, we restricted data acquisition to autosomes in those studies identifying a sex chromosome. In a few studies, a subset of chromosomes had to be ignored because they showed massive macro-assembly problems (large marker gaps, or genetic map position failing to increase monotonically over large chromosome regions). All raw data sets are available on the Dryad repository (https://doi.org/10.5061/dryad.p1j7n43).

2.2 | Characterizing the broad-scale distribution of CO across taxa

A first goal was to visualize the broad-scale distribution of CO along chromosomes across all species within each of the three organismal kingdoms (animals, plants, fungi). For studies providing cM vs. Mb data, this initially required calculating the CO rate (cM/Mb) for intervals of adjacent markers. To achieve comparability, physical midpoint positions of marker intervals were then scaled according to a standard chromosome length of one, and CO rates were divided by their respective chromosome average rate (i.e., mean-standardized, Houle, 1992; qualitatively similar results leading to the same conclusions were obtained by standardizing CO rates by the chromosome-specific standard deviation, or by performing no standardization at all). These adjustments made variation in CO rate within chromosomes independent from differences in physical length and in absolute CO rate among chromosomes and organisms. Within each species, we next combined standardized CO rates from all chromosomes according to their relative chromosome position. For this, we assigned CO rate data points from all chromosomes (scaled to unit length) to one of 25 adjacent windows and computed for each window the median CO rate across chromosomes (using the mean to combine the data points within an organism produced similar result). Finally, the species-specific CO rates thus summarized were averaged across species within each kingdom for each of the 25 chromosome windows for visualization (data available as Appendix S2). We also calculated 95% confidence intervals (CIs) around the window-specific means by bootstrap resampling among the species 10,000 times (Manly, 2007; throughout the study, CIs around point estimates were calculated analogously). For selected species, we also visualized the standardized CO rate along an exemplary chromosome at the original marker resolution and physical chromosome scale.

2.3 | Influence of the centromere on the broad-scale distribution of CO

The above analysis revealed a general broad-scale reduction in CO rate across the centres of chromosomes (see Section 3). To gain insights into potential underlying causes, we explored to what extent this heterogeneity in CO rate is related to the position of the centromere, a chromosome region essential for proper chromosome segregation and exhibiting a reduced CO rate (Talbert & Henikoff, 2010). This analysis focused on the subset of species for which centromere positions were available. These positions had to be inferred from DNA sequence motifs or other physical markers, not from the distribution of CO. We further ignored species with short chromosomes (less than ~20 Mb on average), because we found that pronounced broad-scale heterogeneity in CO rate was often lacking on short chromosomes (see Section 3), thus precluding a meaningful analysis of the centromere's role in driving such heterogeneity. In the 17 total species satisfying these criteria (Table 1; references to the studies characterizing centromere position in these species are given in Table S1), we assigned all chromosomes to one of six total morphological categories. These included metacentric, submetacentric, subtelocentric, acrocentric and telocentric chromosomes, as defined by a decreasing ratio of the short to the long chromosome arm (Levan, Fredga, & Sandberg, 1964). These five categories thus provide a crude description of how central or peripheral the centromere is located within a chromosome. The sixth category was the holocentric chromosomes lacking a single well-defined centromere. Here the spindle fibres guiding chromosome segregation can attach along the entire chromosome (Dernburg, 2001; Melters, Paliulis, Korf, & Chan, 2012). To assess whether the broad-scale reduction in CO rate across chromosome centres is determined by centromere position, we took two qualitative, visual approaches (quantitative analysis was precluded by heterogeneity in the quality of centromere position information across studies): first, we focused on species with the same morphology across all chromosomes. For these species, we graphed the median standardized CO rate for each of the 25 chromosome windows as described above and then compared the distribution of the CO rate between species differing in chromosome morphology. The second approach focused on different chromosome morphologies occurring *within* species. We here again plotted window-specific standardized CO rates, but this time separately for each chromosome morphology category within a species (at least three chromosomes per morphological category were required). In both analyses, our prediction was that if the centromere position determines the broad-scale CO landscape, chromosomes exhibiting peripheral centromeres should lack a systematic reduction in CO rate around chromosome centres.

2.4 | Relationship between CO rate and chromosome length

Observations during data acquisition raised the possibility that the strength of the reduction in CO rate within chromosome centres relative to peripheries (see Section 3) could be related to chromosome length. This idea was investigated both among and within species. For the former, we reused the standardized CO rates calculated for each chromosome in each species as described above. For each chromosome, we calculated the mean CO rate across all

TABLE 1 Species of higher eukaryotes included in our meta-analysis of crossover rate, sorted by organismal kingdom and class (animals) or family (plants)

Kingdom	Class/Family	Species	Common name	Author
Animals	Actinopterygii	*Colossoma macropomum*	Tambaqui	Nunes et al. (2017)
Animals	Actinopterygii	*Cyprinus carpio*	Carp	Xu et al. (2014)
Animals	Actinopterygii	*Danio rerio*	Zebrafish	Bradley et al. (2011)
Animals	Actinopterygii	*Gasterosteus aculeatus*[w,c]	Threespine stickleback	Roesti et al. (2013)
Animals	Actinopterygii	*Ictalurus punctatus*	Catfish	Liu et al. (2016)
Animals	Actinopterygii	*Lates calcarifer*[w]	Asian seabass	Wang et al. (2017)
Animals	Aves	*Ficedula albicollis*[w]	Collared flycatcher	Kawakami et al. (2014)
Animals	Aves	*Gallus gallus*	Chicken	Groenen et al. (2009)
Animals	Aves	*Taeniopygia guttata*	Zebra finch	Backström et al. (2010)
Animals	Branchiopoda	*Daphnia magna*[w]	Daphnia	Dukić, Berner, Roesti, Haag, and Ebert (2016)
Animals	Chromadorea	*Caenorhabditis briggsae*[c]	Nematode	Ross et al. (2011)
Animals	Chromadorea	*Caenorhabditis elegans*[c]	Nematode	Rockman and Kruglyak (2009)
Animals	Insecta	*Aedes aegypti*[w,c]	Yellow fever mosquito	Juneja et al. (2014)
Animals	Insecta	*Apis mellifera*	Honeybee	Solignac et al. (2007)
Animals	Insecta	*Bactrocera cucurbitae*	Melon fly	Sim and Geib (2017)
Animals	Insecta	*Bombus terrestris*[w]	Bumblebee	Liu et al. (2017)
Animals	Insecta	*Drosophila melanogaster*	Fruit fly	Comeron, Ratnappan, and Bailin (2012)
Animals	Insecta	*Heliconius melpomene*[w,c]	Postman butterfly	Davey et al. (2016)
Animals	Insecta	*Laupala kohalensis x paranigra*[w]	Cricket	Blankers, Oh, Bombarely, and Shaw (2017)
Animals	Insecta	*Nasonia vitripennis*[w,c]	Wasp	Niehuis et al. (2010)
Animals	Mammalia	*Bos taurus*[c]	Cattle	Arias, Keehan, Fisher, Coppieters, and Spelman (2009)
Animals	Mammalia	*Canis lupus familiaris*[c]	Dog	Wong et al. (2010)
Animals	Mammalia	*Cervus elaphus*[w,c]	Red deer	Johnston et al. (2017)
Animals	Mammalia	*Felis catus*[c]	Cat	Li et al. (2016)
Animals	Mammalia	*Homo sapiens*[w,c]	Human	Jensen-Seaman et al. (2004)
Animals	Mammalia	*Mus musculus*[c]	Mouse	Jensen-Seaman et al. (2004)
Animals	Mammalia	*Ovis aries*[c]	Sheep	Johnston et al. (2016)
Animals	Mammalia	*Pan troglodytes verus*[w,c]	Chimpanzee	Auton et al. (2012)
Animals	Mammalia	*Rattus norvegicus*[c]	Rat	Jensen-Seaman et al. (2004)
Animals	Mammalia	*Sus scrofa*	Pig	Tortereau et al. (2012)
Fungi	Dothideomycetes	*Zymoseptoria tritici*[w]		Croll, Lendenmann, Stewart, and McDonald (2015)
Fungi	Saccharomycetes	*Saccharomycetes cerevisiae*	Baker's yeast	Cherry et al. (2012)
Fungi	Sordariomycetes	*Fusarium graminearum*[w]		Laurent et al. (2017)
Plants	Amaranthaceae	*Beta vulgaris*[c]	Sugar beet	Dohm et al. (2014)
Plants	Asteraceae	*Helianthus annuus*	Sunflower	Renaut et al. (2013)
Plants	Brassicaceae	*Arabidopsis thaliana*	Thale cress	Giraut et al. (2011)
Plants	Brassicaceae	*Brassica napus*	Rapeseed	Wang et al. (2015b)
Plants	Brassicaceae	*Brassica rapa*	Chinese cabbage	Huang, Yang, Zhang, and Cao (2017)
Plants	Cucurbitaceae	*Citrullus lanatus*	Watermelon	Ren et al. (2012)
Plants	Cucurbitaceae	*Cucumis melo*	Melon	Argyris et al. (2015)
Plants	Fabaceae	*Cicer arietinum*	Chickpea	Deokar et al. (2014)
Plants	Fabaceae	*Glycine max*	Soybean	Schmutz et al. (2010)
Plants	Fabaceae	*Phaseolus vulgaris*[c]	Common bean	Bhakta, Jones, and Vallejos (2015)
Plants	Juglandaceae	*Juglans regia*[w]	Walnut	Luo et al. (2015)
Plants	Malvaceae	*Gossypium hirsutum*	Cotton	Wang et al. (2015a)

(Continues)

TABLE 1 (Continued)

Kingdom	Class/Family	Species	Common name	Author
Plants	Malvaceae	*Theobroma cacao*	Cocoa	Argout et al. (2011)
Plants	Phrymaceae	*Mimulus guttatus*[w]	Monkey flower	Holeski et al. (2014)
Plants	Poaceae	*Brachypodium distachyon*	Purple false brome	Huo et al. (2011)
Plants	Poaceae	*Oryza sativa*	Rice	Tian et al. (2009)
Plants	Poaceae	*Setaria italica*	Foxtail millet	Zhang et al. (2012)
Plants	Poaceae	*Sorghum bicolor*	Sorghum	Bekele, Wieckhorst, Friedt, and Snowdon (2013)
Plants	Poaceae	*Triticum aestivum*	Wheat	Gardner, Wittern, and Mackay (2016)
Plants	Poaceae	*Zea mays*	Maize	Bauer et al. (2013)
Plants	Rosaceae	*Fragaria vesca*	Woodland strawberry	Shulaev et al. (2011)
Plants	Rosaceae	*Malus pumila*	Apple	Daccord et al. (2017)
Plants	Rosaceae	*Prunus persica*	Peach	International Peach Genome Initiative (2013)
Plants	Rutaceae	*Citrus clementina*	Clementine	Wu et al. (2014)
Plants	Salicaceae	*Populus deltoides*	Eastern cottonwood	Tong et al. (2016)
Plants	Salicaceae	*Populus simonii*	Simon poplar	Tong et al. (2016)
Plants	Solanaceae	*Capsicum annuum*	Pepper	Hill et al. (2015)
Plants	Solanaceae	*Solanum lycopersicum*	Tomato	Tomato Genome Consortium (2012)
Plants	Solanaceae	*Solanum tuberosum*	Potato	Endelman and Jansky (2016)

Superscripts following species names indicate studies in which the crosses or pedigrees underlying genetic mapping were derived from wild individuals (*w*), and for which information on centromere position was available (*c*).

marker intervals having their physical mid-point within 10 Mb from either chromosome tip (using 5 Mb only produced very similar results) and divided this value by the chromosome-wide average CO rate. The resulting "CO periphery-bias" provided a standardized descriptor of the CO distribution along a chromosome, with a value near one indicating a relatively evenly distributed CO rate, and greater positive values indicating a concentration of CO towards the chromosome tips. Next, we defined the length of each chromosome as the Mb position of the terminal marker interval mid-point, calculated mean CO periphery-bias and chromosome length across the chromosomes within each species and assessed if chromosome length predicted the CO distribution when using species as data points. This was carried out using Spearman's rank correlation (hereafter simply "correlation" because we always applied the Spearman method to quantify the strength of association between variables) and included all species except the single one lacking physical chromosome positions (Nunes et al., 2017). The correlation between CO periphery-bias and chromosome length was further explored *within* species (i.e., using chromosomes as data points). To ensure sufficient sensitivity, this latter analysis was restricted to species represented by at least six chromosomes in our data set, exhibiting at least one chromosome longer than 30 Mb, and showing an at least twofold length difference between the shortest and longest chromosome. The distribution of species-specific correlation coefficients was then evaluated within animals ($N = 16$) and plants ($N = 11$) separately (the species used for this analysis are listed in Appendix S3). To confirm the adequacy of our CO periphery-bias metric, we repeated the above analyses by quantifying the distribution of CO along a chromosome using two alternative methods: the coefficient of a quadratic regression of standardized CO rate vs. Mb position and the ratio of mean peripheral to central CO rate based on the crude centre-periphery delimitation used in Berner and Roesti (2017). All these analyses produced qualitatively similar results supporting the same conclusions, so we report only results obtained with the main method (data available as Appendix S3).

Because the above analysis indicated that the CO distribution within chromosomes was related to chromosome length, we next explored whether chromosome length also predicted the average chromosome-wide CO rate (i.e., cM/Mb across the entire chromosome, ignoring *within*-chromosome heterogeneity). Again, this analysis was performed among and within species (data available as Appendix S4). For the former, we cumulated genetic and physical map length across all chromosomes of each species in our data set for which raw cM information was available ($N = 52$). Dividing total cumulative genetic map length by its physical counterpart then yielded an estimate of the average CO rate for a chromosome—and of the average CO rate across the entire genome—in a given species. Finally, we examined if this quantity was related to median chromosome length when using species as data points. In an analogous analysis within species, we divided genetic by physical map length for each chromosome and calculated the correlation between this average CO rate and physical length across chromosomes within each species represented by at least six chromosomes in our data set. The distribution of correlation coefficients was then evaluated across species separately within each kingdom.

2.5 | Relationship between CO rate and gene density

A major evolutionary consequence of CO is that selectively relevant genetic variation from multiple copies of a given chromosome can be recombined. The efficacy of this process depends on the distribution of CO relative to the distribution of genetic information units along chromosomes. Our finding of heterogeneity in the distribution of CO thus raised the important question whether the density of genes is also heterogeneous at the scale of entire chromosomes. To explore this question, we first retrieved data from BIOMART (http://www.biomart.org) on the physical location of protein-coding genes along chromosomes (considering only autosomes, when known) in all species with annotated genomes (16 animals, 14 plants, 3 fungi; total $N = 33$). The broad-scale distribution of gene density was then characterized analogously to the distribution of CO along chromosomes: each chromosome in each species was scaled to unit length and divided into 25 windows of equal width, and the number of genes falling into each window was determined. Variation in gene density among species, and among chromosomes within species, was accounted for by scaling window-specific gene counts along a given chromosome by the mean number of genes across all windows on that chromosome. Relative gene density thus obtained was then summarized for each species by calculating the median value over all chromosomes for each of the 25 windows. Finally, we averaged the species-specific relative gene densities for each window and estimated the associated 95% bootstrap CIs, separately for each kingdom. In addition, we quantified the strength of the association between gene density and CO rate within each animal and plant species by the correlation coefficient calculated with window-specific median values as data points, and evaluated the distribution of this statistic in both kingdoms (due to small sample size, this distribution was again not evaluated in fungi). We note that these analyses made the assumption that the density of potential selective targets in a chromosome region can be expressed based on gene counts. This assumption appears reasonable, given a strong correspondence between gene number and total coding sequence length at least at a broad scale (Berner & Roesti, 2017).

2.6 | Relationship between CO rate and the magnitude of population differentiation

In a final set of analyses, we examined how the interaction between broad-scale heterogeneity in CO rate and divergent natural selection can influence patterns of genetic differentiation in genome-wide marker-based population comparisons. We here reused single nucleotide polymorphism (SNP) data generated through RAD sequencing in threespine stickleback fish (*Gasterosteus aculeatus*) adapted to ecologically different habitats (ocean, lake, stream) in the Vancouver Island region (Canada; Roesti, Gavrilets, Hendry, Salzburger, & Berner, 2014; Roesti, Hendry, Salzburger, & Berner, 2012). Specifically, we focused on a pair of populations that diverged between the lake and its adjacent outlet stream habitat in the Boot Lake watershed (our "lake–stream" population comparison), and a pair involving a marine and a geographically close freshwater (stream-resident) population (Sayward estuary and Robert's stream; our "marine–freshwater" population comparison). Detailed information on the ecology and adaptive divergence of these populations and on the generation of the SNP data is provided in Berner, Adams, Grandchamp, and Hendry (2008), Berner, Grandchamp, and Hendry (2009) and Roesti et al. (2012, 2014). For both population comparisons, SNPs were first quality filtered as described in Roesti et al. (2014) and then used to calculate the absolute allele frequency difference (AFD) as a simple metric of population differentiation (Shriver et al., 1997). Considering data from the 20 autosomes only (i.e., the known sex chromosome was excluded) and using only the one SNP per RADtag producing the highest AFD, we obtained differentiation values from 3,622 SNPs for the lake–stream and 9,351 SNPs for the marine–freshwater comparison. Given a genome size of ~460 Mb for threespine stickleback (Jones et al., 2012), the marker resolution in these data sets was relatively low (the expected spacing between SNPs was ~130 and 50 kb) but still sufficient to characterize broad-scale trends in population differentiation (Roesti et al., 2012, 2014).

We first generated differentiation profiles along chromosomes for each population comparison, averaging AFD values from individual SNPs across nonoverlapping sliding windows of 1 Mb. Next, we assessed to what extent differentiation values were correlated between the two—ecologically different (lake–stream vs. marine–freshwater)—population comparisons. For this, we calculated the correlation between the two comparisons across all nonoverlapping, genome-wide sliding windows, considering different window sizes: 0.1, 0.2, 0.5, 1, 2, 3 and 4 Mb. As this analysis revealed an increasingly strong correlation between the two differentiation profiles with increasing window size (see Section 3), we hypothesized that increasing window size should also lead to a stronger genome-wide association between CO rate and differentiation within each population comparison. We tested this prediction by calculating the average CO rate for all windows based on genome-wide CO rate data from Roesti, Moser, and Berner (2013) and quantified how strongly this variable was correlated with the average population differentiation calculated for the same windows. As above, this procedure was repeated for different window sizes ranging from 0.1 to 4 Mb. Finally, our observation that heterogeneity in the distribution of CO is related to their physical length (see Section 3) motivated two analyses focusing on the relationship between chromosome length and the magnitude of population differentiation. In the first analysis, we calculated for each of the two population comparisons the chromosome-specific overall magnitude of genetic differentiation based on the median AFD value across all SNPs on a chromosome. Then, we calculated the correlation between overall differentiation and chromosome length separately for each population comparison. In the second analysis, we defined the SNPs from the top 5% of the genome-wide AFD distribution in each population comparison as "high-differentiation SNPs" and calculated for each chromosome the proportion of high-differentiation SNPs among the total SNPs on that chromosome (thus accounting for different absolute SNP numbers among chromosomes). Then, we tested if this proportion

was correlated to chromosome length. All these analyses excluded the sex chromosome (19), and additionally chromosome 21; the latter because this chromosome harbours a large (>2 Mb) inversion (Jones et al., 2012; Roesti, Kueng, Moser, & Berner, 2015) confounding the broad-scale CO distribution (including chromosome 21 did not qualitatively change any conclusion). All analyses and plotting were performed with R (R Core Team 2017); codes are available upon request.

3 | RESULTS AND DISCUSSION

3.1 | Data set for meta-analysis

Our literature search identified 62 species in which CO rates were linked to chromosome-level genome assemblies, including 30 animals, 29 plants and 3 fungi (Table 1). Our data set is thus well suited for generalizations about the CO landscape in animals and plants, but less so in the fungal kingdom. The data set is clearly dominated by species of economic relevance and laboratory model systems, which is not surprising, given that generating a chromosome-level genome assembly remains a substantial investment. For the vast majority of species (>90%), CO information suitable to this study was available in graphical form only. To facilitate future investigations, we encourage authors to publish raw genetic map positions in cM together with physical Mb positions for all markers in tabulated and hence more easily accessible form.

3.2 | Reduced CO rate in chromosome centres is a major trend in eukaryotes

Our meta-analysis revealed a striking broad-scale pattern across the animal and plant data sets: chromosome centres displayed a dramatically reduced CO rate compared to the chromosome peripheries (Figure 1). In animals, the rate of peripheral CO was more than 2.5 times higher than the CO rate in the central region of chromosomes, and in plants, this difference was more than fivefold. Animals further displayed a clear drop in CO rate towards the very tips of the chromosomes. Additional exploration of the plant data (including filtering for those species with the highest marker resolution and with annotated and hence probably high-quality genomes, and considering different chromosome length classes; details not presented) strongly suggested that the absence of a (strong) terminal drop in CO rate in plants is real, and not an artefact. In contrast to animals and plants, fungal species did not exhibit a clear broad-scale trend in the distribution of CO, although the data for this organismal kingdom were sparse. To ensure that the pattern seen in animals and plants was not driven by specific taxonomic groups, we additionally analysed data separately for all animal classes and plant families listed in Table 1, provided they were represented by at least three different genera (i.e., ray-finned fishes [Actinopterygii], birds, insects and mammals; Fabaceae, Poaceae and Rosaceae). This confirmed that a reduced CO rate in chromosome centres is taxonomically widespread within the animal and plant kingdoms (Figure S1 in Appendix S1). In addition, we examined if there was an influence of artificial selection on the distribution of CO. The motivation was that strong selection and small population size—typical conditions under domestication—are expected theoretically to impose indirect selection on genetic variants that increase the CO rate (Barton & Otto, 2005), a prediction with mixed empirical support (Burt & Bell, 1987; Muñoz-Fuentes et al., 2015; Rees & Dale, 1974; Ross-Ibarra, 2004). While we see no reason why domestication should drive consistent evolution in the *physical location* of CO along chromosomes, we nevertheless graphed the average CO landscape for the pool of all

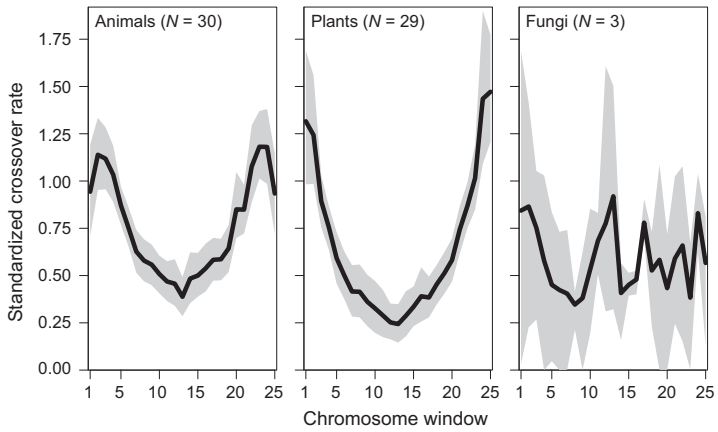


FIGURE 1 Broad-scale heterogeneity in crossover (CO) rate along chromosomes in eukaryotes. Black lines show the average CO rate across all species within each organismal kingdom (sample sizes in parentheses) along chromosomes divided into 25 windows. Associated 95% bootstrap confidence bands are shaded in grey (in fungi with low sample size, this band reflects the data range). To achieve comparability among study systems, all chromosomes were first scaled to unit length and CO rates were mean standardized across windows within each chromosome, and then, the median value per window was calculated across all chromosomes within each species

animals classified as wild (N = 12, Table 1; a meaningful analogous analysis in plants was precluded by the low number of wild species, N = 2). Wild animals also exhibited the strong reduction in CO rate in chromosome centres observed across the complete data sets (Figure S1), ruling out domestication as an explanation for the observed trend in the distribution of CO.

3.3 | Broad-scale heterogeneity in CO rate is not well explained by centromere position

An intuitive explanation for CO to occur primarily towards the peripheries of a chromosome is that CO may be inhibited in the chromosome's centres if this region harbours the centromere. The centromere is a chromosome region typically characterized by a core of DNA sequence repeats serving as the assembly site of the kinetochore, a protein complex to which the spindle fibres required for proper chromosome segregation attach. In addition, the centromere is possibly also involved in chromosome sorting during the very early stages of meiosis (Allshire & Karpen, 2008; Da Ines, Gallego, & White, 2014; Malik & Henikoff, 2009; McFarlane & Humphrey, 2010; Zickler & Kleckner, 2016). Around the centromere, CO is well known to be suppressed (Beadle, 1932; Harushima et al., 1998; Lambie & Roeder, 1986; Mahtani & Willard, 1998; Rahn & Solari, 1986; Sherman & Stack, 1995). In yeast, for instance, molecular components of the kinetochore complex inhibit DNA double-strand breaks—a necessary precursor of CO—near the centromere and prevent DNA breaks in the broader neighbourhood of the centromere to be repaired as CO (Ellermeier et al., 2010; Vincenten et al., 2015).

Two aspects of centromeres, however, challenge their general importance as determinants of the broad-scale chromosomal distribution of CO across species. The first is the centromeres' relatively small size. Consequently, centromere-associated CO suppression may be a relatively localized phenomenon within a chromosome only. Indeed, CO inhibition extends over just a few kilobases around the centromere in budding yeast (Vincenten et al., 2015), and over 2.3 Mb on a rice chromosome investigated (Yan et al., 2005). It is thus not evident how an extensive low-CO region on a chromosome hundreds of megabases in length (see below) could be mediated by the centromere alone. The second aspect challenging the idea that regions of low CO rate in chromosome centres are driven by centromeres is that centromeres are not necessarily located in the physical centre of chromosomes. Hence, if the centromere was a major broad-scale determinant of the CO distribution, we would expect bias in CO rate towards chromosome peripheries to be restricted to chromosomes harbouring the centromeres near their centre. We assessed this prediction qualitatively by comparing the distribution of CO among species with different overall chromosome morphologies, as defined by their relative centromere position. This analysis revealed that species exhibiting exclusively acro- or telocentric chromosomes—that is, having centromeres located close to one chromosome end—still display reduced CO rates across the chromosome centre (or the centre of the longer chromosome arm) (Figure 2, left column; the pattern in species with metacentric chromosomes is shown in Figure S2). Moreover, some species with holocentric chromosomes, hence lacking a single well-defined centromeric domain, show the same broad-scale trend. Similar insights emerged from the comparison of different chromosome morphologies within species (Figure 2, right column; Figure S2). Collectively, these observations in no way challenge that the centromere influences the CO landscape, but show that the centromere alone fails to provide a universal explanation for the general broad-scale reduction in CO rate in chromosome centres seen across taxa.

3.4 | The distribution of CO is predicted by chromosome length

As a next step, we explored if the broad-scale distribution of CO was related to the length of chromosomes. For this, we quantified the relative elevation in CO rate in the chromosome peripheries by our CO periphery-bias statistic, and related this statistic to chromosome length. Pooling all species as data points in a single analysis revealed a clear association: organism lacking a marked reduction in the CO rate in chromosome centres (i.e., exhibiting CO periphery-bias around one) were those displaying short chromosomes, and the CO distribution became increasingly periphery-biased as chromosome length increased (Figure 3a; Spearman's rank correlation: r_S = 0.86, 95% CI: 0.74–0.93). This association also held when analysing animals and plants separately (animals: r_S = 0.79, 95% CI: 0.52–0.93; plants: r_S = 0.87, 95% CI: 0.68–0.95). A clear relationship between chromosome length and CO periphery-bias also emerged *within* species: the correlation between these two variables among chromosomes was almost consistently positive, and often strongly so, in both animals and plants (Figure 3b).

In combination, these analyses make clear that the magnitude of periphery-bias in CO rate is a function of the length of a chromosome. Organisms lacking a pronounced reduction in CO rate in chromosome centres are those having short chromosomes, typically below some 20 Mb. This includes species such as *Arabidopsis thaliana*, some social insects (honeybee, bumblebee) and, importantly, all fungi in our data set (the CO distribution along a representative chromosome from each of three species with short chromosomes is shown in Figure S3, left). Fungi are known to generally have short chromosomes (Cervelatti, Ferreira-Nozawa, Aquino-Ferreira, Fachin, & Martinez-Rossi, 2004), and this may well be the simple reason why our analysis of this group indicates a CO distribution qualitatively different from that seen in the other kingdoms (Figure 1). By contrast, the species in our data set exhibiting very long chromosomes, including wheat, maize, pepper, sunflower and several mammals, generally have CO restricted to short peripheral chromosome regions separated by a vast CO desert (three examples are shown in Figure S3, right). Based on these observations, it is tempting to propose a simple conceptual model in which CO occurs preferentially within a characteristic distance from the chromosome tips, and the total length of a chromosome then determines the physical extent of the central low-CO region (Figure 4). As suggested by Figure 3a, this characteristic distance may often be within some 10 Mb (see also

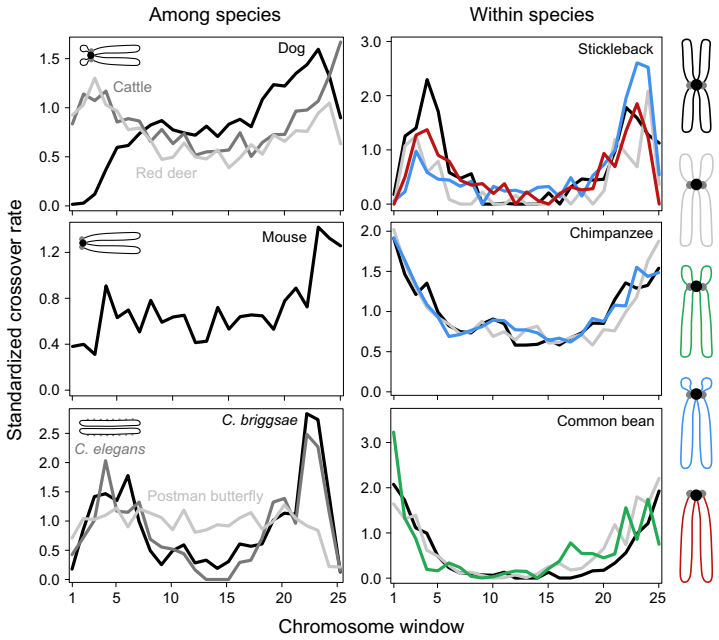


FIGURE 2 Centromere position does not explain the widely observed broad-scale reduction in CO rate around the centre of chromosomes. The left column shows average CO rate profiles for species with uniform chromosome morphologies. From top to bottom, these are acrocentric, telocentric and holocentric chromosomes. The right column shows CO rate profiles for three exemplary species with variable chromosome morphologies. The colour coding refers to the chromosome morphology schematics, ordered from top to bottom by increasingly peripheral centromere position: metacentric in black, submetacentric in grey, subtelocentric in green, acrocentric in blue and telocentric in red. In each species, the profiles represent averages across at least three chromosomes for a given morphology. For both averaging and plotting, chromosomes were always oriented such that the shorter chromosome arm was on the left. In the chromosome schematics, the constriction separating the chromosome arms is shown in black and the kinetochore(s) in grey. CO rates and chromosome lengths were standardized for comparability as in Figure 1. Additional evidence from further species is presented in Figure S2

Johnston, Berenos, Slate, & Pemberton, 2016; Pratto et al., 2014; Roesti et al., 2013; Smeds, Mugal, Qvarnstrom, & Ellegren, 2016). According to this view, short chromosomes consist primarily of high-CO periphery. Before we evaluate the plausibility of this model in the light of evidence from investigations of the mechanisms governing meiosis, we consider a prediction regarding the genome-wide CO rate implicit in this model.

3.5 | Peripheral CO causes a negative association between average CO rate and chromosome length

The above conceptual model predicts that genomes consisting of short chromosomes, and hence mainly peripheral chromosome regions exhibiting a high CO rate, should show higher overall (i.e., genome-wide) CO rates than genomes consisting of long chromosomes with physically extensive centres of low CO rate. This prediction was confirmed: among species, we found a striking negative, nonlinear association between the average CO rate of chromosomes (or, equivalently, cumulative cM/Mb across the entire genome) and median chromosome length (Figure 5a, left panel, all species pooled; $r_S = -0.92$, 95% CI: −0.95 to −0.83). Extreme CO rates occurred in the species with the smallest chromosomes, including the two fungus species available for this specific analysis. A similar relationship emerged when analysing animals ($r_S = -0.90$, 95% CI: −0.96 to −0.77) and plants ($r_S = -0.84$, 95% CI: −0.94 to −0.60) separately. Interestingly, this relationship could be approximated by making the simplified assumption of a universal genetic map length of 50 cM per chromosome, corresponding to a single CO per chromosome and meiosis, and dividing this standard genetic map length by different physical chromosome lengths covering the range of median chromosome lengths observed in our organisms (Figure 5a, right panel). Chromosome length thus emerges as a remarkably strong predictor of the genome-wide CO rate among species, challenging the recent suggestion (Stapley, Feulner, Johnston, Santure, & Smadja, 2017) that features of genome architecture are relatively unimportant determinants of broad-scale CO rate variation among eukaryotes. Our insights from the analysis among species were further reinforced by relating CO rate to chromosome length *within* species. In both animals and plants, the correlation between these two variables was generally strongly negative among chromosomes (Figure 5b; the

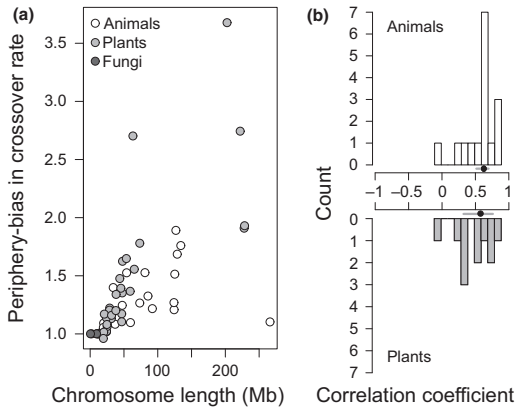


FIGURE 3 The strength to which the CO rate is biased towards the chromosome peripheries is related to chromosome length. (a) Strength of periphery-bias in CO rate plotted against chromosome length, using species means as data points (colour-coded by kingdom). (b) Distribution of the strength of the correlation between CO periphery-bias and chromosome length within animal and plant species (N = 16 and 11). A positive coefficient indicates that in a given species, the CO rate becomes more strongly biased towards the chromosome tips as chromosome length increases. The dots and error bars next to the X-axis show the median correlation coefficient and the associated 95% bootstrap CI for each kingdom

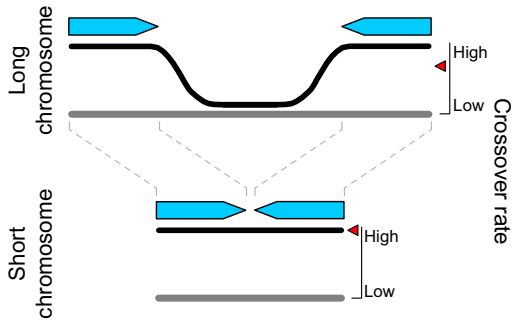


FIGURE 4 A simple conceptual model for the broad-scale CO distribution in which the probability of CO along chromosomes (grey bars) is high only within a characteristic distance from the chromosome tips (blue arrows). On a long chromosome, these regions are separated, thus producing a region of low CO rate (black curve) in the chromosome's centre. On a short chromosome, by contrast, the probability of CO and hence the observed CO distribution is relatively homogeneous. Consequently, the average CO rate of a chromosome (red triangles) is a function of its length

distribution of correlation coefficients was not visualized for fungi because only two species were available, but both species also showed a negative coefficient; see also Backström et al., 2010; Giraut et al., 2011; International Human Genome Sequencing Consortium 2001; Jensen-Seaman et al., 2004; Johnston, Huisman, Ellis, & Pemberton, 2017; Kaback, Guacci, Barber, & Mahon, 1992; Roesti et al., 2013; Smeds et al., 2016; Tortereau et al., 2012).

Taken together, these analyses suggest that the genome-wide CO rate in eukaryotes is strongly determined by the relative proportion of the genome having a high rate of CO, that is, the proportion of peripheral DNA. For a given genome size, an organism may thus achieve a higher rate of CO—and thus stronger reshuffling of genetic variation—by distributing its total DNA among a greater number of smaller chromosomes. In the animal kingdom, particularly high genome-wide CO rates have been reported from social hymenopteran insects (Sirviö et al., 2006; Wilfert, Gadau, & Schmid-Hempel, 2007), with 37 cM/Mb in the honey bee (Beye et al., 2006; Liu et al., 2015; Solignac, Mougel, Vautrin, Monnerot, & Cornuet, 2007), 14 cM/Mb in *Pogonomyrmex* ants (Sirviö, Pamilo, Johnson, Page, & Gadau, 2011a), 9.7 cM/Mb in the common wasp (Sirviö, Johnston, Wenseleers, & Pamilo, 2011b) and 8.7 cM/Mb in the bumblebee (Liu et al., 2017). The evolutionary reason for this high average CO rate is not well understood, but perhaps reflects the need for rapid adaptation to fast-evolving pathogens to which social insects seem particularly strongly exposed, or for compensating the sex-limited recombination associated with haplo-diploid sex determination (Sirviö et al., 2006; Wilfert et al., 2007). However, these CO rates do not appear exceptionally high when taking heterogeneity in CO along chromosomes into account: species exhibiting a very low genome-wide CO rate (e.g., sunflower, wheat: 0.3 and 1.1 cM/Mb) reach similarly high CO rates as social insects when averaging exclusively over the terminal 5 Mb on either side of each chromosome (14.9 and 9.3 cM/Mb; see also Roesti et al., 2013; Pratto et al., 2014)—that is, when considering a total chromosome segment approximating median chromosome length in the honeybee (10.7 Mb) or bumblebee (14.5 Mb). Hence, a key feature of CO distinguishing some social insect species from other animals is that their genomes are split into many short chromosomes (Wilfert et al., 2007) lacking extensive central regions with a low CO rate. The same likely applies to fungi, a group also exhibiting very high genome-wide recombination rates and short chromosomes (Awadalla, 2003; Cervelatti et al., 2004; Stapley et al., 2017; Wilfert et al., 2007). Like social insects, many fungi also interact with other organisms as pathogens or through symbiosis, and have limited opportunity for recombination due to extensive haploid life phases, both of which may have selected for a high CO rate across their genomes. These considerations highlight the limited information conveyed by estimates of the average, genome-wide CO rate. Understanding to which extent genetic variation is shuffled by CO requires knowledge about the actual distribution of the CO rate within and among chromosomes.

3.6 | What causes the high CO rate in chromosome peripheries?

We have argued that a conceptual model in which CO happens mainly within some distance from the chromosome tips, irrespective of total chromosome length, helps explain associations between the average CO rate, the distribution of CO and chromosome length. Is

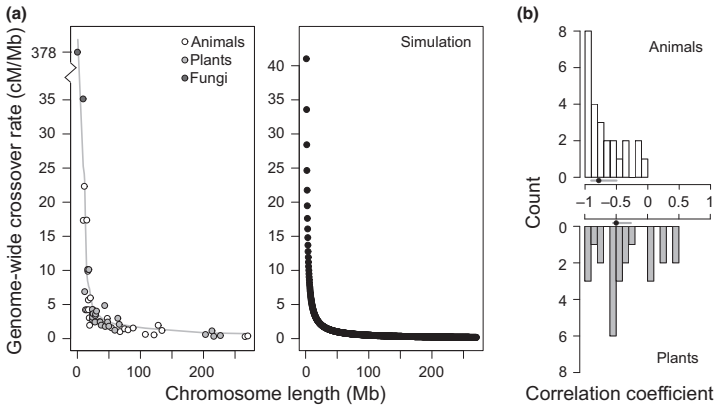


FIGURE 5 The average CO rate is related to chromosome length. (a) The left panel displays the average genome-wide CO rate against median chromosome length, using species as data points (colour-coded by kingdom; only species for which genetic map length was available were considered, $N = 52$). The light grey curve shows a nonparametric fit (loess; moving average with bandwidth of 0.5) to the data pooled across the organismal kingdoms. The *Y*-axis is broken to allow for an extreme value. The right panel shows the corresponding relationship as simulated across the empirically observed chromosome length range based on the assumption of a universal single CO per chromosome and meiosis (50 cM genetic map length). The empirical chromosome length range was sampled at 1000 evenly spaced points. (b) Distribution of the strength of the correlation between the chromosome-wide CO rate and chromosome length within animal and plant species ($N = 25$ in both kingdoms). A negative correlation coefficient indicates that in a given species, long chromosomes exhibit a lower CO rate than short chromosomes. The dots and error bars next to the *X*-axis show the median correlation coefficient and the associated 95% bootstrap CI for each kingdom

there any mechanistic evidence in support of such a model? Indeed, an elegant explanation for periphery-bias in CO rate is related to the choreography of chromosomes and the spatio-temporal sequence of recombination initiation during meiosis. Higher eukaryotes generally share a phase in the early stages of meiosis during which the telomeres (i.e., the chromosome tips) aggregate at the nuclear membrane, while the chromosome centres remain closer to the nucleus' centre (Harper, Golubovskaya, & Cande, 2004; Naranjo & Corredor, 2008; Scherthan et al., 1996; Zickler & Kleckner, 2016). This stage, often referred to as the "meiotic bouquet" (Scherthan, 2001), is followed by rapid chromosome oscillations during which the chromosomes alternately disperse and aggregate (Klutstein & Cooper, 2014). This movement is again coordinated by the telomeres, which remain in contact with the nuclear membrane. The function of the bouquet and the oscillations remains incompletely understood, but very likely they enable homology search and the pairing of chromosomes (Bass et al., 2000; Chacon, Delivani, & Tolic, 2016; Curtis, Lukaszewski, & Chrzastek, 1991; Ding, Yamamoto, Haraguchi, & Hiraoka, 2004; Gerton & Hawley, 2005; Lee, Conrad, & Dresser, 2012; Lefrancois, Rockmill, Xie, Roeder, & Snyder, 2016; Page & Hawley, 2003). Intriguingly, these telomere-guided processes may also influence the location of CO along chromosomes: evidence from several organisms suggests that synapsis, that is, the establishment of a physical connection between homologous chromosomes, and associated DNA double-strand breaks required for CO are initiated from the chromosome tips, and that the repair of these breaks as CO is more likely in the chromosome peripheries than the centres (Anderson & Stack, 2005; Bass et al., 2000; Brown et al., 2005; Croft & Jones, 1989; Higgins, Osman, Jones, & Franklin, 2014; Klutstein & Cooper, 2014; Lukaszewski, 1997; Pratto et al., 2014; Viera, Santos, & Rufas, 2009; Xiang, Miller, Ross, Alvarado, & Hawley, 2014). The telomere-guided initiation of chromosome homology search and recombination could thus be part of the explanation why CO occurs primarily towards the chromosome peripheries (Scherthan et al., 1996; Zickler & Kleckner, 2016).

Another potentially important aspect is crossover interference, that is, the inhibition of additional CO in the vicinity of an existing CO along a chromosome (Muller, 1916; Sturtevant, 1915). This is suggested by sexual dimorphism in the distribution of CO: remarkably consistently across species, the enrichment of CO near the telomeres is more pronounced in the male than the female sex (Broman, Murray, Sheffield, White, & Weber, 1998; Cox et al., 2009; Giraut et al., 2011; Johnston et al., 2016, 2017; Lien et al., 2011; Ma et al., 2015; Smeds et al., 2016). Interestingly, the sexes also appear to differ in the structural organization of meiotic chromosomes, with the paired homologous chromosomes being less condensed in oocytes than spermatocytes (Tease & Hulten, 2004). If CO interference operates at the same spatial (i.e., μm, not base pairs) scale in both sexes, CO interference will therefore extend over a shorter base pair distance in females than males (Kochakpour & Moens, 2008; Petkov, Broman, Szatkiewicz, & Paigen, 2007). Consequently, male CO may be strongly limited to the chromosome tips where the first obligate CO occurs, whereas in females, additional CO may occur along the chromosomes, thus leading to a more homogeneous distribution of CO and an elevated overall CO count in females. Evaluating these ideas will require a more complete mechanistic understanding of meiosis based on experimental evidence from a wide variety of organismal systems.

3.7 | Implications of broad-scale heterogeneity in CO rate for evolutionary genomics

So far, we have demonstrated predictable broad-scale heterogeneity in CO rate along chromosomes, but what is the significance of this variation to evolutionary genomic theory and empirical analysis? A pivotal aspect of the CO rate is that it determines the physical scale of linked selection within a genome: allele frequency shifts driven by natural selection on a given locus extend relatively deeply into the locus' nonselected chromosomal neighbourhood when the locus is situated in a low-CO region, but decay over a shorter physical scale when the locus resides in a high-CO region (Maynard Smith & Haigh, 1974). Such linked selection has received distinct names in different evolutionary contexts, including "background selection" when the selected polymorphisms arise from new deleterious mutation (Charlesworth, Morgan, & Charlesworth, 1993; Hudson & Kaplan, 1995; Nordborg, Charlesworth, & Charlesworth, 1996); "genetic draft" when the polymorphisms arise from new beneficial mutations (Gillespie, 2000); and "gene flow barrier" when the polymorphisms arise from genetic exchange between populations under divergent selection (Aeschbacher, Selby, Willis, & Coop, 2017; Barton, 1979; Barton & Bengtsson, 1986; Berner & Roesti, 2017; Feder & Nosil, 2010; Roesti et al., 2014). These processes differ in detail. For instance, background selection is considered inevitable and ubiquitous because the majority of mutations are generally considered deleterious (Lynch et al., 1999). However, although plausibly occurring more rarely, new *beneficial* alleles arising from mutation will rise from initially low frequency, causing more intense selection than low-frequency deleterious mutations (Cutter & Payseur, 2013). Also, both background selection and genetic draft rely on new mutations and therefore have little impact on short time scales (Burri, 2017). By contrast, gene flow barriers can emerge rapidly by selection on standing genetic variation, although they require some level of genetic exchange between diverging populations (Berner & Roesti, 2017; see also Samuk et al., 2017). Despite these nuances, the different forms of linked selection can be housed under a single conceptual roof because they are all similarly affected by the CO rate. Importantly, natural selection implies a reduction in effective population size and hence elevated stochasticity in the transmission of genetic variation across generations (genetic drift) at a locus. By modifying the physical scale of linked selection around a locus, the CO rate thus influences the strength of drift in a genome region, and hence, the level of genetic diversity maintained within populations and of genetic differentiation among populations (Charlesworth, 1998; Cutter & Payseur, 2013; Nachman & Payseur, 2012).

Combined with the widespread broad-scale reduction in CO rate in chromosome centres relatively to chromosome peripheries, the above theory on linked selection predicts that populations should commonly harbour relatively low levels of genetic variation in chromosome centres, and that comparisons between populations should find relatively elevated genetic differentiation in chromosome centres (Figure 6). Genome-wide marker-based studies indeed support this prediction (Burri et al., 2015; Carneiro et al., 2014; Dutoit et al., 2017; Gante et al., 2016; Roesti et al., 2012, 2013; Samuk et al., 2017; Tine et al., 2014). However, elucidating the details in the underlying linked selection will often be difficult. The reason is that background selection and genetic draft are notoriously hard to disentangle (Comeron, 2017; Cutter & Payseur, 2013). Moreover, gene flow between population and species can persist over long time spans (Berner & Salzburger, 2015), so that selection on new mutations and selection against immigrant alleles (gene flow barrier) may shape patterns in genetic variation jointly (Aeschbacher et al., 2017). Only when divergence between populations is so recent that a substantial contribution from selection on new mutations can be ruled out, broad-scale patterns in genetic diversity and population differentiation can be ascribed to linked selection caused by heterogeneity along chromosomes in the strength of gene flow barriers.

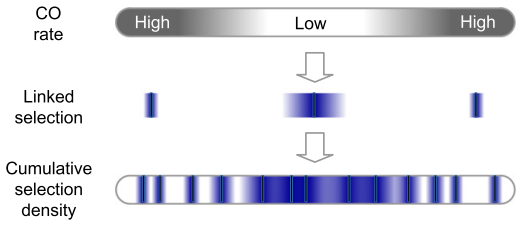


FIGURE 6 Relationship between heterogeneous CO rate and selection density along a chromosome. If the CO rate is reduced in the chromosome centre relative to the peripheries (top), selection on a locus (shown as black vertical bar) in the centre will cause linked selection to extend deeper into the locus' chromosomal neighbourhood than in the peripheries (middle; the strength of linked selection is visualized by the blue shade). Consequently, selection at many loci—due to continued mutation over long timescales and/or to gene flow between populations in selectively different habitats—will generate a relatively elevated cumulative density of linked selection in the chromosome centre (bottom). This elevated selection density implies a reduction in effective population size, and hence stronger drift, in chromosome centres. Chromosome centres will therefore harbour less genetic variation within populations and exhibit elevated genetic differentiation among populations, relative to the peripheries

The above reflections make clear that heterogeneity in the distribution of CO across the genome is a key determinant of heterogeneity in the distribution of genetic variation within and between populations. Equally important, however, is the distribution of selective targets along chromosomes: if regions of low CO rate coincide with regions of low gene density, selection on new mutations or maladaptive immigrant alleles may not necessarily drive heterogeneity in diversity and differentiation across the genome (Aeschbacher et al., 2017; Cutter & Payseur, 2013; Payseur & Nachman, 2002). The reason is that the wider physical extent of linked selection in a low-CO region is counterbalanced by a reduced probability of selection to target this region in the first place. Understanding how heterogeneous CO rate modifies the consequences of selection across the genome thus benefits from knowledge about the broad-scale distribution of selection targets along chromosomes. This motivated our analysis of the density of genes along chromosomes, considering the subset of species in our data set for which annotated

genomes were available. We found no indication of systematic broad-scale heterogeneity in gene density along chromosomes in animals or fungi (Figure 7a): in these groups, genes appear distributed relatively evenly along chromosomes (noting that sample size for fungi was low). In striking contrast, a clear pattern emerged in plants: on average, gene density proved ~3.5 times higher towards the chromosome peripheries than in the chromosome centres. These findings were confirmed by examining the correlations between gene density and CO rate within each species: in animals, the correlation coefficients peaked around zero, whereas in plants, the correlations were consistently positive and mostly very strong (Figure 7b). Our investigation thus highlights a peculiarity of plant genomes: genes tend to be located in chromosome regions crossing over relatively frequently (see also Gaut, Wright, Rizzon, Dvorak, & Anderson, 2007; Mezard, 2006; Schnable, Hsia, & Nikolau, 1998). As recombination is a potent mechanism of DNA loss counteracting the proliferation of transposable elements, it is possible that in many plant species, chromosome centres with a low CO rate have developed into gene-poor regions through the accumulation of repetitive DNA (Bennetzen, 2000; Puchta, 2005; Hawkins, Grover, & Wendel, 2008; Schubert & Vu, 2016; see also Nam & Ellegren, 2012; Kapusta, Suh, & Feschotte, 2017). Nevertheless, the heterogeneity in CO rate across plant genomes on average still exceeds the heterogeneity in gene density, although not as strongly as in animals (dotted lines in Figure 7a). The consequences of natural selection should thus tend to be more profound in chromosome centres than in the peripheries in both taxonomic groups, but particularly strongly so in animals.

3.8 | Empirical demonstration of analytical challenges of broad-scale heterogeneity in CO rate to evolutionary genomics

As described above, a relatively reduced CO rate across chromosome centres in combination with selection can drive systematically elevated population differentiation in chromosome centres. This has serious but insufficiently recognized implications to analytical approaches commonly employed in evolutionary genomics. Importantly, the identification of so-called outlier loci—that is, genetic markers showing particularly strong population differentiation relative to the genome-wide background level and hence considered footprints of divergent selection—can be misleading when using outlier detection approaches ignoring heterogeneity in the CO landscape. Such outliers will tend to be overrepresented in genome regions of low CO rate (Noor & Bennett 2009; Berner & Roesti, 2017) because loci under selection and their selectively neutral chromosomal neighbourhood can reach stronger population differentiation through cumulative linked selection in low-CO regions (Roesti et al., 2012, 2013; Aeschbacher et al., 2017; see Roesti et al., 2012 for a pragmatic approach to adjust marker data for such broad-scale heterogeneity in differentiation). A related inferential problem can arise in investigations of genomic parallelism in evolution. The extent to which repeated adaptive phenotypic divergence in multiple population pairs occurs by responses to divergent selection in the same genes is an important question in evolutionary genomics (Arendt & Reznick, 2008; Bailey, Blanquart, Bataillon, & Kassen, 2017). Popular approaches to addressing this question include evaluating the

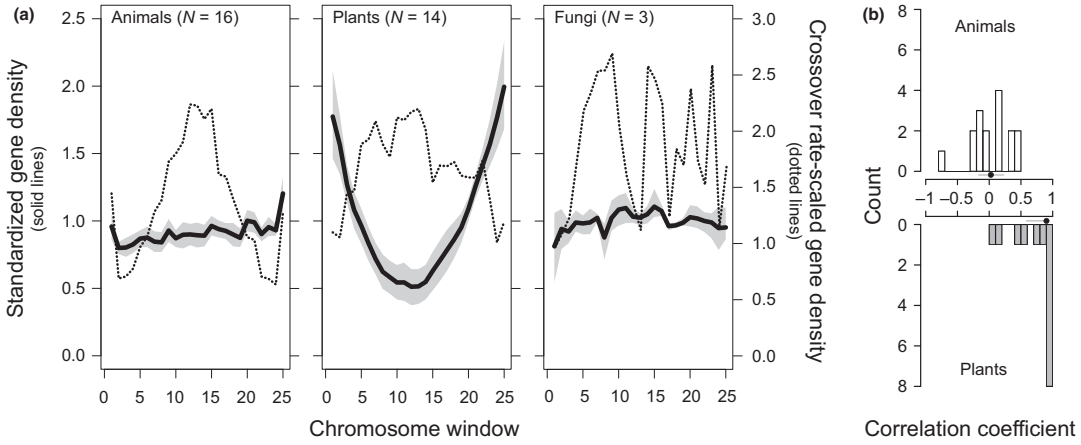


FIGURE 7 (a) Broad-scale distribution of gene density along chromosomes in each kingdom (sample sizes in parentheses). Following the conventions from Figure 1, each chromosome was scaled to unit length and divided into 25 windows, gene density was standardized by the average value across windows for each chromosome, and window-specific median values across all chromosomes were calculated within each species. The solid black lines, referring to the left *Y*-axis scale, are mean values across species, with 95% bootstrap CIs shown as grey bands. The dotted lines, referring to the right *Y*-axis scale, represent standardized gene density divided by standardized CO rate (see Figure 1) for each window and hence express the density of genes along chromosomes relative to the CO rate. (b) Distribution of the coefficients of correlation between gene density and CO rate within species, shown separately for animals and plants. The correlations were calculated based on chromosome window-specific median values for both variables. The dots and error bars next to the X-axis show the median correlation coefficient and the associated 95% bootstrap CI for each kingdom

proportion of high-differentiation outliers (individual markers, or chromosome windows) shared among multiple population comparisons, or to examine whether a correlation in the magnitude of differentiation in markers or chromosome windows exists among multiple population comparisons. Shared outliers and/or correlated differentiation are then often interpreted as indication that divergent natural selection has targeted the same genes in multiple population pairs, and hence as evidence of parallel evolution at the molecular level. However, such analyses are frequently performed with low physical marker resolution (recent examples: Egger, Roesti, Böhne, Roth, & Salzburger, 2017; Perreault-Payette et al., 2017; Ravinet et al., 2016; Raeymaekers et al., 2017; Rougemont et al., 2017; Stuart et al., 2017; Trucchi, Frajman, Haverkamp, Schönswetter, & Paun, 2017). Consequently, single markers are highly unlikely to coincide with polymorphisms under direct selection. Shared population differentiation captured by such marker or chromosome window data may thus primarily mirror common patterns in cumulative linked selection density shaped by a shared broad-scale CO landscape, thus precluding reliable conclusions about (non)parallelism in the specific targets of divergent selection (Berner & Roesti, 2017). This view is particularly plausible when shared patterns in genome-wide differentiation emerge across lineages separated for a long time (Burri et al., 2015; Dutoit et al., 2017; Renaut, Owens, & Rieseberg, 2014; Van Doren et al., 2017; see also Hobolth, Dutheil, Hawks, Schierup, & Mailund, 2011). We emphasize that studies using high-density markers (e.g., as obtained by whole-genome sequencing) are not immune to such confounding if marker-specific differentiation data are averaged across large chromosome windows.

To illustrate these conceptual issues with empirical data, we re-analysed relatively low-resolution SNP data from two ecologically distinct population comparisons of threespine stickleback fish (Roesti et al., 2012, 2014), that is, a lake–stream and a marine–freshwater population pair. Chromosome window-based profiles of population differentiation revealed strikingly elevated differentiation in chromosome centres, a pattern evident in *both ecologically different* population comparisons (Figure 8a). As a consequence, the magnitude of window-specific differentiation was correlated between the two

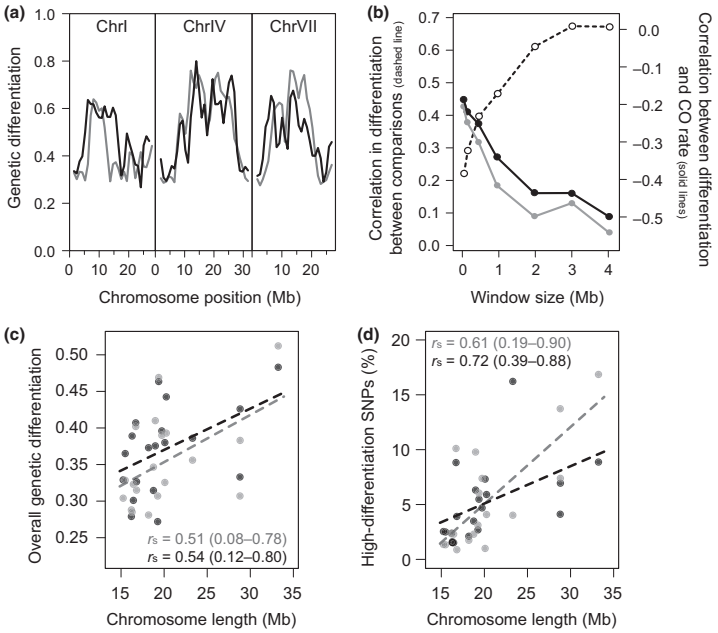


FIGURE 8 (a) Genetic differentiation (quantified by the absolute allele frequency difference, AFD) in a lake–stream (black) and a marine–freshwater (grey; same colour coding used throughout the graphic) stickleback population comparison along the three largest chromosomes exhibiting a particularly pronounced reduction in CO rate around their centres (Roesti et al., 2013). The profiles show mean differentiation across all SNPs for nonoverlapping chromosome windows of 1 Mb. (b) The strength of the correlation between the lake–stream and the marine–freshwater population comparison in the magnitude of differentiation across chromosome windows increases with increasing window size (0.1–4 Mb), shown by the dashed line referring to the left *Y*-axis. Likewise, the negative association between average population differentiation and CO rate across chromosome windows increases within each population comparison as window size increases (solid lines, referring to the right *Y*-axis). (c) and (d) illustrate that both the magnitude of overall differentiation (median AFD) and the relative proportion of SNPs displaying high differentiation (upper 5% of the genome-wide AFD distribution) are correlated to the length of chromosomes in each population comparison (Spearman correlations and their 95% bootstrap CIs are given in each box). Note that the proportion of high-differentiation SNPs is adjusted for the total number of SNPs along a focal chromosome, and hence, a high value indicates a relative excess of strongly differentiated SNPs on a chromosome

population pairs, increasingly strongly so when averaging differentiation values across SNPs for increasingly large physical windows (Figure 8b). A naive interpretation of this association would be that selection has targeted the same genes in both population comparisons. A more parsimonious explanation, however, is that irrespective of the precise targets of selection and the underlying ecological context in each population pair, gene flow barriers have driven shared patterns of broad-scale neutral differentiation across the genome. Indeed, stickleback exhibit strongly reduced CO rates in chromosome centres (Roesti et al., 2013; Glazer, Killingbeck, Mitros, Rokhsar, & Miller, 2015; see also Figure 2), and adaptive divergence in both lake–stream and marine–freshwater stickleback systems is well known to occur in the face of gene flow and to involve selection on a large number of genes (Berner et al., 2009; Hagen, 1967; Jones, Brown, Pemberton, & Braithwaite, 2006; Jones et al., 2012; Lescak et al., 2015; Roesti et al., 2014, 2015; Terekhanova et al., 2014)—conditions facilitating the emergence of heterogeneous differentiation through variation in the strength of gene flow barriers along chromosomes (Berner & Roesti, 2017). (Note that divergence in both population pairs is postglacial and hence too recent for mutation-based linked selection to significantly influence differentiation profiles.) Accordingly, both population comparisons also exhibited a negative genome-wide correlation between population differentiation and CO rate, a relationship increasing in strength with decreasing analytical resolution (Figure 8b). Moreover, in line with the general observation that the relative reduction in CO rate around chromosome centres increases with chromosome length, we found stronger overall population differentiation and a relative excess of high-differentiation SNPs (i.e., adjusted for total SNP number on a chromosome) on longer chromosomes (Figure 8c, d). In diverging stickleback populations, chromosome length thus appears to influence genome-wide heterogeneity in the opportunity for genetic exchange between populations by determining heterogeneity in the strength of gene flow barriers along chromosomes.

The above analytical challenges emphasize the value of two resources in evolutionary genomics: the first is a chromosome-level genome assembly. Combined with genetic map data, an assembly allows characterizing the CO landscape and recognizing broad-scale trends in diversity and differentiation, thus potentially revealing an interaction between the distribution of CO and selection density. The second key resource is a high marker resolution. Inference about targets of selection—a major goal in many evolutionary genomic studies—requires an analytical resolution much finer than the broad scale of the patterns in genetic variation driven by heterogeneity in CO-mediated selection density. We argue that in the light of widespread variation in CO rate along chromosomes, these two aspects deserve more weight when designing evolutionary genomic investigations.

4 | CONCLUSION

Our synthesis of the distribution of crossovers in 62 species reveals a taxonomically widespread trend for CO to occur primarily towards the peripheries of chromosomes. This distribution of CO rate is closely linked to the physical length of chromosomes and strongly influences the genome-wide average CO rate. Although we can rule out the centromere as major driver of this chromosome-scale heterogeneity in CO rate, substantial progress in recombination research will be needed to identify the underlying mechanistic determinants, and to allow assessing to what extent these determinants are shared among organisms. Given the strong impact of the CO landscape on the consequences of natural selection to genetic diversity within and between populations, quantifying and embracing heterogeneity in CO rate should become a standard element of analytical approaches and their interpretation in evolutionary genomics.

ACKNOWLEDGEMENTS

We kindly thank all researchers who have made this synthesis possible by making primary data available; Scott Hawley and Danny Miller for discussion; Gleb Ebert for helping extract data; Reto Burri and two anonymous reviewers for valuable feedback; Louis Bernatchez for inviting this review. DB was supported financially by the Swiss National Science Foundation (SNF; grant 31003A_165826) and the University of Basel. MR was supported by the SNF and a Janggen-Pöhn fellowship.

CONFLICT OF INTEREST

The authors declare no competing financial interests.

DATA ACCESSIBILITY

Table S1 and Figures S1–S3 can be found online in Appendix S1. Supporting data are provided as Appendices S2, S3 and S4. The raw data sets are available on Dryad (https://doi.org/10.5061/dryad.p1j7n43).

AUTHOR CONTRIBUTIONS

Q.H. contributed to literature search, led data extraction, and performed data analysis; T.G.L. contributed to study design, literature search and data extraction, performed data analysis and contributed to manuscript writing; M.R. contributed to study design, literature search and data analysis; D.B. designed and supervised the study, contributed to literature search and data extraction, developed analytical tools, performed data analysis and wrote the paper with feedback from all co-authors.

ORCID

Telma G. Laurentino http://orcid.org/0000-0001-7879-5251
Marius Roesti http://orcid.org/0000-0002-7408-4804

REFERENCES

Aeschbacher, S., Selby, J. P., Willis, J. H., & Coop, G. (2017). Population–genomic inference of the strength and timing of selection against gene flow. *Proceedings of the National Academy of Sciences of the*

United States of America, *114*, 7061–7066. https://doi.org/10.1073/pnas.1616755114

Akhunov, E. D., Goodyear, A. W., Geng, S., Qi, L.-L., Echalier, B., Gill, B. S., … Dvorak, J. (2003). The organization and rate of evolution of wheat genomes are correlated with recombination rates along chromosome arms. *Genome Research*, *13*, 753–763. https://doi.org/10.1101/gr.808603

Allshire, R. C., & Karpen, G. H. (2008). Epigenetic regulation of centromeric chromatin: Old dogs, new tricks? *Nature Review Genetics*, *9*, 923–937. https://doi.org/10.1038/nrg2466

Anderson, L. K., & Stack, S. M. (2005). Recombination nodules in plants. *Cytogenetic and Genome Research*, *109*, 198–204. https://doi.org/10.1159/000082400

Arendt, J., & Reznick, D. (2008). Convergence and parallelism reconsidered: What have we learned about the genetics of adaptation? *Trends in Ecology and Evolution*, *23*, 26–32. https://doi.org/10.1016/j.tree.2007.09.011

Argout, X., Salse, J., Aury, J. M., Guiltinan, M. J., Droc, G., Gouzy, J., … Lanaud, C. (2011). The genome of *Theobroma cacao*. *Nature Genetics*, *43*, 101–108. https://doi.org/10.1038/ng.736

Argyris, J. M., Ruiz-Herrera, A., Madriz-Masis, P., Sanseverino, W., Morata, J., Pujol, M., … Garcia-Mas, J. (2015). Use of targeted SNP selection for an improved anchoring of the melon (*Cucumis melo* L.) scaffold genome assembly. *BMC Genomics*, *16*, 4. https://doi.org/10.1186/s12864-014-1196-3

Arias, J. A., Keehan, M., Fisher, P., Coppieters, W., & Spelman, R. (2009). A high density linkage map of the bovine genome. *BMC Genetics*, *10*, 18. https://doi.org/10.1186/1471-2156-10-18

Auton, A., Fledel-Alon, A., Pfeifer, S., Venn, O., Segurel, L., Street, T., … McVean, G. (2012). A fine-scale chimpanzee genetic map from population sequencing. *Science*, *336*, 193–198. https://doi.org/10.1126/science.1216872

Awadalla, P. (2003). The evolutionary genomics of pathogen recombination. *Nature Review Genetics*, *4*, 50–60. https://doi.org/10.1038/nrg964

Backström, N., Forstmeier, W., Schielzeth, H., Mellenius, H., Nam, K., Bolund, E., … Ellegren, H. (2010). The recombination landscape of the zebra finch *Taeniopygia guttata* genome. *Genome Research*, *20*, 485–495. https://doi.org/10.1101/gr.101410.109

Bailey, S. F., Blanquart, F., Bataillon, T., & Kassen, R. (2017). What drives parallel evolution? How population size and mutational variation contribute to repeated evolution. *BioEssays*, *39*, 1–9.

Barton, N. H. (1979). Gene flow past a cline. *Heredity*, *43*, 333–339. https://doi.org/10.1038/hdy.1979.86

Barton, N. H. (1995). A general model for the evolution of recombination. *Genetics Research*, *65*, 123–144. https://doi.org/10.1017/S0016672300033140

Barton, N., & Bengtsson, B. O. (1986). The barrier to genetic exchange between hybridizing populations. *Heredity*, *57*, 357–376. https://doi.org/10.1038/hdy.1986.135

Barton, N. H., & Otto, S. P. (2005). Evolution of recombination due to random drift. *Genetics*, *169*, 2353–2370. https://doi.org/10.1534/genetics.104.032821

Bass, H. W., Riera-Lizarazu, O., Ananiev, E. V., Bordoli, S. J., Rines, H. W., Phillips, R. L., … Cande, W. Z. (2000). Evidence for the coincident initiation of homolog pairing and synapsis during the telomere-clustering (bouquet) stage of meiotic prophase. *Journal of Cell Science*, *113*, 1033–1042.

Baudat, F., Imai, Y., & de Massy, B. (2013). Meiotic recombination in mammals: Localization and regulation. *Nature Review Genetics*, *14*, 794–806. https://doi.org/10.1038/nrg3573

Bauer, E., Falque, M., Walter, H., Bauland, C., Camisan, C., Campo, L., … Schön, C.-C. (2013). Intraspecific variation of recombination rate in maize. *Genome Biology*, *14*, 1–17.

Beadle, G. W. (1932). A possible influence of the spindle fibre on crossing-over in *Drosophila*. *Proceedings of the National Academy of Sciences of the United States of America*, *18*, 160–165. https://doi.org/10.1073/pnas.18.2.160

Bekele, W. A., Wieckhorst, S., Friedt, W., & Snowdon, R. J. (2013). High-throughput genomics in sorghum: From whole-genome resequencing to a SNP screening array. *Plant Biotechnology Journal*, *11*, 1112–1125. https://doi.org/10.1111/pbi.12106

Bennetzen, J. L. (2000). Transposable element contributions to plant gene and genome evolution. *Plant Molecular Biology*, *42*, 251–269. https://doi.org/10.1023/A:1006344508454

Berner, D., Adams, D. C., Grandchamp, A.-C., & Hendry, A. P. (2008). Natural selection drives patterns of lake-stream divergence in stickleback foraging morphology. *Journal of Evolutionary Biology*, *21*, 1653–1665. https://doi.org/10.1111/j.1420-9101.2008.01583.x

Berner, D., Grandchamp, A.-C., & Hendry, A. P. (2009). Variable progress toward ecological speciation in parapatry: Stickleback across eight lake-stream transitions. *Evolution*, *63*, 1740–1753. https://doi.org/10.1111/j.1558-5646.2009.00665.x

Berner, D., & Roesti, M. (2017). Genomics of adaptive divergence with chromosome-scale heterogeneity in crossover rate. *Molecular Ecology*, *26*, 6351–6369. https://doi.org/10.1111/mec.14373

Berner, D., & Salzburger, W. (2015). The genomics of organismal diversification illuminated by adaptive radiations. *Trends in Genetics*, *31*, 491–499. https://doi.org/10.1016/j.tig.2015.07.002

Beye, M., Gattermeier, I., Hasselmann, M., Gempe, T., Schioett, M., Baines, J. F., … Page, R. E. (2006). Exceptionally high levels of recombination across the honey bee genome. *Genome Research*, *16*, 1339–1344. https://doi.org/10.1101/gr.5680406

Bhakta, M. S., Jones, V. A., & Vallejos, C. E. (2015). Punctuated distribution of recombination hotspots and demarcation of pericentromeric regions in *Phaseolus vulgaris* L. *PLoS ONE*, *10*, e0116822. https://doi.org/10.1371/journal.pone.0116822

Blankers, T., Oh, K. P., Bombarely, A., & Shaw, K. L. (2017). The genomic architecture of a rapid island radiation: Mapping chromosomal rearrangements and recombination rate variation in *Laupala*. *bioRxiv*. https://doi.org/10.1101/160952

Bradley, K. M., Breyer, J. P., Melville, D. B., Broman, K. W., Knapik, E. W., & Smith, J. R. (2011). An SNP-based linkage map for zebrafish reveals sex determination loci. *G3: Genes, Genomes, Genetics*, *1*, 3–9. https://doi.org/10.1534/g3.111.000190

Broman, K. W., Murray, J. C., Sheffield, V. C., White, R. L., & Weber, J. L. (1998). Comprehensive human genetic maps: Individual and sex-specific variation in recombination. *The American Journal of Human Genetics*, *63*, 861–869. https://doi.org/10.1086/302011

Brown, P. W., Judis, L. A., Chan, E. R., Schwartz, S., Seftel, A., Thomas, A., & Hassold, T. J. (2005). Meiotic synapsis proceeds from a limited number of subtelomeric sites in the human male. *American Journal of Human Genetics*, *77*, 556–566. https://doi.org/10.1086/468188

Burri, R. (2017). Interpreting differentiation landscapes in the light of long-term linked selection. *Evolution Letters*, *1*, 118–131. https://doi.org/10.1002/evl3.14

Burri, R., Nater, A., Kawakami, T., Mugal, C. F., Olason, P. I., Smeds, L., … Ellegren, H. (2015). Linked selection and recombination rate variation drive the evolution of the genomic landscape of differentiation across the speciation continuum of *Ficedula* flycatchers. *Genome Research*, *25*, 1656–1665. https://doi.org/10.1101/gr.196485.115

Burt, A. (2000). Sex, recombination, & the efficacy of selection – was Weismann right? *Evolution*, *54*, 337–351.

Burt, A., & Bell, G. (1987). Mammalian chiasma frequencies as a test of two theories of recombination. *Nature*, *326*, 803–805.

Carneiro, M., Albert, F. W., Afonso, S., Pereira, R. J., Burbano, H., Campos, R., … Ferrand, N. (2014). The genomic architecture of population divergence between subspecies of the European rabbit. *PLoS Genetics*, *10*, e1003519. https://doi.org/10.1371/journal.pgen.1003519

Cervelatti, E. P., Ferreira-Nozawa, M. S., Aquino-Ferreira, R., Fachin, A. L., & Martinez-Rossi, N. M. (2004). Electrophoretic molecular karyotype of the dermatophyte *Trichophyton rubrum*. *Genetics and Molecular Biology*, *27*, 99–102. https://doi.org/10.1590/S1415-47572004000100016

Chacon, M. R., Delivani, P., & Tolic, I. M. (2016). Meiotic nuclear oscillations are necessary to avoid excessive chromosome associations. *Cell Reports*, *17*, 1632–1645. https://doi.org/10.1016/j.celrep.2016.10.014

Charlesworth, B. (1998). Measures of divergence between populations and the effect of forces that reduce variability. *Molecular Biology and Evolution*, *15*, 538–543. https://doi.org/10.1093/oxfordjournals.molbev.a025953

Charlesworth, B., Morgan, M. T., & Charlesworth, D. (1993). The effect of deleterious mutations on neutral molecular variation. *Genetics*, *134*, 1289–1303.

Cherry, J. M., Hong, E. L., Amundsen, C., Balakrishnan, R., Binkley, G., Chan, E. T., … Wong, E. D. (2012). *Saccharomyces* genome database: The genomics resource of budding yeast. *Nucleic Acids Research*, *40*, D700–D705. https://doi.org/10.1093/nar/gkr1029

Choi, K., & Henderson, I. R. (2015). Meiotic recombination hotspots – a comparative view. *The Plant Journal*, *83*, 52–61. https://doi.org/10.1111/tpj.12870

Comeron, J. M. (2017). Background selection as null hypothesis in population genomics: Insights and challenges from *Drosophila* studies. *Proceedings of the Royal Society B*, *372*, 20160471.

Comeron, J. M., Ratnappan, R., & Bailin, S. (2012). The many landscapes of recombination in *Drosophila melanogaster*. *PLoS Genetics*, *8*, e1002905. https://doi.org/10.1371/journal.pgen.1002905

Cox, A., Ackert-Bicknell, C. L., Dumont, B. L., Ding, Y., Bell, J. T., Brockmann, G. A., … Broman, K. A. (2009). A new standard genetic map for the laboratory mouse. *Genetics*, *182*, 1335–1344. https://doi.org/10.1534/genetics.109.105486

Croft, J. A., & Jones, G. H. (1989). Meiosis in *Mesostoma ehrenbergii ehrenbergii*. IV. Recombination nodules in spermatocytes and a test of the correspondence of late recombination nodules and chiasmata. *Genetics*, *121*, 255–262.

Croll, D., Lendenmann, M. H., Stewart, E., & McDonald, B. A. (2015). The impact of recombination hotspots on genome evolution of a fungal plant pathogen. *Genetics*, *201*, 1213–1228. https://doi.org/10.1534/genetics.115.180968

Curtis, C. A., Lukaszewski, A. J., & Chrzastek, M. (1991). Metaphase I pairing of deficient chromosomes and genetic mapping of deficiency breakpoints in common wheat. *Genome*, *34*, 553–560. https://doi.org/10.1139/g91-085

Cutter, A. D., & Payseur, B. A. (2013). Genomic signatures of selection at linked sites: Unifying the disparity among species. *Nature Review Genetics*, *14*, 262–274. https://doi.org/10.1038/nrg3425

Da Ines, O., Gallego, M. E., & White, C. I. (2014). Recombination-independent mechanisms and pairing of homologous chromosomes during meiosis in plants. *Molecular Plant*, *7*, 492–501. https://doi.org/10.1093/mp/sst172

Daccord, N., Celton, J. M., Linsmith, G., Becker, C., Choisne, N., Schijlen, E., … Bucher, E. (2017). High-quality de novo assembly of the apple genome and methylome dynamics of early fruit development. *Nature Genetics*, *49*, 1099–1106. https://doi.org/10.1038/ng.3886

Davey, J. W., Chouteau, M., Barker, S. L., Maroja, L., Baxter, S. W., Simpson, F., … Jiggins, C. D. (2016). Major improvements to the *Heliconius melpomene* genome assembly used to confirm 10 chromosome fusion events in 6 million years of butterfly evolution. *G3: Genes, Genomes, Genetics*, *6*, 695–708. https://doi.org/10.1534/g3.115.023655

Deokar, A. A., Ramsay, L., Sharpe, A. G., Diapari, M., Sindhu, A., Bett, K., … Tar'an, B. (2014). Genome wide SNP identification in chickpea for use in development of a high density genetic map and improvement of chickpea reference genome assembly. *BMC Genomics*, *15*, 708. https://doi.org/10.1186/1471-2164-15-708

Dernburg, A. F. (2001). Here, there, & everywhere: Kinetochore function on holocentric chromosomes. *Journal of Cell Biology*, *153*, F33–F38. https://doi.org/10.1083/jcb.153.6.F33

Ding, D. Q., Yamamoto, A., Haraguchi, T., & Hiraoka, Y. (2004). Dynamics of homologous chromosome pairing during meiotic prophase in fission yeast. *Developmental Cell*, *6*, 329–341. https://doi.org/10.1016/S1534-5807(04)00059-0

Dohm, J. C., Lange, C., Holtgrawe, D., Sorensen, T. R., Borchardt, D., Schulz, B., … Himmelbauer, H. (2012). Palaeohexaploid ancestry for Caryophyllales inferred from extensive gene-based physical and genetic mapping of the sugar beet genome (*Beta vulgaris*). *The Plant Journal*, *70*, 528–540. https://doi.org/10.1111/j.1365-313X.2011.04898.x

Dohm, J. C., Minoche, A. E., Holtgrawe, D., Capella-Gutierrez, S., Zakrzewski, F., Tafer, H., … Himmelbauer, H. (2014). The genome of the recently domesticated crop plant sugar beet (*Beta vulgaris*). *Nature*, *505*, 546–549. https://doi.org/10.1038/nature12817

Dukić, M., Berner, D., Roesti, M., Haag, C. R., & Ebert, D. (2016). A high-density genetic map reveals variation in recombination rate across the genome of *Daphnia magna*. *BMC Genetics*, *17*, 137.

Dutoit, L., Vijay, N., Mugal, C. F., Bossu, C. M., Burri, R., Wolf, J. B. W., & Ellegren, H. (2017). Covariation in levels of nucleotide diversity in homologous regions of the avian genome long after completion of lineage sorting. *Proceedings of the Royal Society B*, *284*, 20162756. https://doi.org/10.1098/rspb.2016.2756

Egger, B., Roesti, M., Böhne, A., Roth, O., & Salzburger, W. (2017). Demography and genome divergence of lake and stream populations of an East African cichlid fish. *Molecular Ecology*, *26*, 5016–5030. https://doi.org/10.1111/mec.14248

Ellermeier, C., Higuchi, E. C., Phadnis, N., Holm, L., Geelhood, J. L., Thon, G., & Smith, G. R. (2010). RNAi and heterochromatin repress centromeric meiotic recombination. *Proceedings of the National Academy of Sciences of the United States of America*, *107*, 8701–8705. https://doi.org/10.1073/pnas.0914160107

Endelman, J. B., & Jansky, S. H. (2016). Genetic mapping with an inbred line-derived F2 population in potato. *Theoretical and Applied Genetics*, *129*, 935–943. https://doi.org/10.1007/s00122-016-2673-7

Feder, J. L., & Nosil, P. (2010). The efficacy of divergence hitchhiking in generating genomic islands during ecological speciation. *Evolution*, *64*, 1729–1747. https://doi.org/10.1111/j.1558-5646.2009.00943.x

Felsenstein, J. (1974). The evolutionary advantage of recombination. *Genetics*, *78*, 737–756.

Fisher, R. A. (1930). *The genetical theory of natural selection*. Oxford, UK: Oxford University. https://doi.org/10.5962/bhl.title.27468

Gante, H. F., Matschiner, M., Malmstrom, M., Jakobsen, K. S., Jentoft, S., & Salzburger, W. (2016). Genomics of speciation and introgression in Princess cichlid fishes from Lake Tanganyika. *Molecular Ecology*, *25*, 6143–6161. https://doi.org/10.1111/mec.13767

Gardner, K. A., Wittern, L. M., & Mackay, I. J. (2016). A highly recombined, high-density, eight-founder wheat MAGIC map reveals extensive segregation distortion and genomic locations of introgression segments. *Plant Biotechnology Journal*, *14*, 1406–1417. https://doi.org/10.1111/pbi.12504

Gaut, B. S., Wright, S. I., Rizzon, C., Dvorak, J., & Anderson, L. K. (2007). Recombination: An underappreciated factor in the evolution of plant genomes. *Nature Review Genetics*, *8*, 77–84. https://doi.org/10.1038/nrg1970

Gerton, J. L., & Hawley, R. S. (2005). Homologous chromosome interactions in meiosis: Diversity amidst conservation. *Nature Review Genetics*, *6*, 477–487. https://doi.org/10.1038/nrg1614

Gillespie, J. H. (2000). Genetic drift in an infinite population: The pseudohitchhiking model. *Genetics*, *155*, 909–919.

Giraut, L., Falque, M., Drouaud, J., Pereira, L., Martin, O. C., & Mezard, C. (2011). Genome-wide crossover distribution in *Arabidopsis thaliana* meiosis reveals sex-specific patterns along chromosomes. *PLoS Genetics*, *7*, e1002354. https://doi.org/10.1371/journal.pgen.1002354

Glazer, A. M., Killingbeck, E. E., Mitros, T., Rokhsar, D. S., & Miller, C. T. (2015). Genome assembly improvement and mapping convergently evolved skeletal traits in sticklebacks with genotyping-by-sequencing. *G3: Genes, Genomes, Genetics*, *5*, 1463–1472. https://doi.org/10.1534/g3.115.017905

Groenen, M. A., Wahlberg, P., Foglio, M., Cheng, H. H., Megens, H. J., Crooijmans, R. P., … Andersson, L. (2009). A high-density SNP-based linkage map of the chicken genome reveals sequence features correlated with recombination rate. *Genome Research*, *19*, 510–519.

Hagen, D. W. (1967). Isolating mechanisms in threespine sticklebacks (*Gasterosteus*). *Journal of the Fisheries Research Board of Canada*, *24*, 1637–1692. https://doi.org/10.1139/f67-138

Harper, L., Golubovskaya, I., & Cande, W. Z. (2004). A bouquet of chromosomes. *Journal of Cell Science*, *117*, 4025–4032. https://doi.org/10.1242/jcs.01363

Hartfield, M., & Otto, S. P. (2011). Recombination and hitchhiking of deleterious alleles. *Evolution*, *65*, 2421–2434. https://doi.org/10.1111/j.1558-5646.2011.01311.x

Harushima, Y., Yano, M., Shomura, P., Sato, M., Shimano, T., Kuboki, Y., … Sasaki, T. (1998). A high–density rice genetic linkage map with 2275 markers using a single F–2 population. *Genetics*, *148*, 479–494.

Hassold, T., & Hunt, P. (2001). To err (meiotically) is human: The genesis of human aneuploidy. *Nature Review Genetics*, *2*, 280–291. https://doi.org/10.1038/35066065

Hawkins, J. S., Grover, C. E., & Wendel, J. F. (2008). Repeated big bangs and the expanding universe: Directionality in plant genome size evolution. *Plant Science*, *174*, 557–562. https://doi.org/10.1016/j.plantsci.2008.03.015

Higgins, J. D., Osman, K., Jones, G. H., & Franklin, F. C. H. (2014). Factors underlying restricted crossover localization in barley meiosis. *Annual Review of Genetics*, *48*, 29–47. https://doi.org/10.1146/annurev-genet-120213-092509

Hill, T., Ashrafi, H., Chin-Wo, S. R., Stoffel, K., Truco, M.-J., Kozik, A., … Van Deynze, A. (2015). Ultra-high density, transcript-based genetic maps of pepper define recombination in the genome and synteny among related species. *G3: Genes, Genomes, Genetics*, *5*, 2341–2355. https://doi.org/10.1534/g3.115.020040

Hill, W. G., & Robertson, A. (1966). Effect of linkage on limits to artificial selection. *Genetics Research*, *8*, 269. https://doi.org/10.1017/S0016672300010156

Hobolth, A., Dutheil, J. Y., Hawks, J., Schierup, M. H., & Mailund, T. (2011). Incomplete lineage sorting patterns among human, chimpanzee, and orangutan suggest recent orangutan speciation and widespread selection. *Genome Research*, *21*, 349–356. https://doi.org/10.1101/gr.114751.110

Holeski, L. M., Monnahan, P., Koseva, B., McCool, N., Lindroth, R. L., & Kelly, J. K. (2014). A high-resolution genetic map of yellow monkeyflower identifies chemical defense QTLs and recombination rate variation. *G3: Genes, Genomes, Genetics*, *4*, 813–821. https://doi.org/10.1534/g3.113.010124

Houle, D. (1992). Comparing evolvability and variability of quantitative traits. *Genetics*, *130*, 195–204.

Huang, L., Yang, Y., Zhang, F., & Cao, J. (2017). A genome-wide SNP-based genetic map and QTL mapping for agronomic traits in Chinese cabbage. *Scientific Reports*, *7*, 46305. https://doi.org/10.1038/srep46305

Hudson, R. R., & Kaplan, N. L. (1995). Deleterious background selection with recombination. *Genetics*, *141*, 1605–1617.

Hunter, N. (2007). Meiotic recombination. In A. Aguilera, & R. Rothstein (Eds.), *Molecular genetics of recombination* (pp. 381–442). Berlin, Germany: Springer. https://doi.org/10.1007/978-3-540-71021-9

Huo, N., Garvin, D. F., You, F. M., McMahon, S., Luo, M. C., Gu, Y. Q., … Vogel, J. P. (2011). Comparison of a high-density genetic linkage map to genome features in the model grass *Brachypodium distachyon*. *Theoretical and Applied Genetics*, *123*, 455–464. https://doi.org/10.1007/s00122-011-1598-4

International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*, 860–921.

International Peach Genome Initiative, Verde, I., Abbott, A. G., Scalabrin, S., Jung, S., Shu, S., … Rokhsar, D. S. (2013). The high-quality draft genome of peach (*Prunus persica*) identifies unique patterns of genetic diversity, domestication and genome evolution. *Nature Genetics*, *45*, 487–494. https://doi.org/10.1038/ng.2586

Jensen-Seaman, M. I., Furey, T. S., Payseur, B. A., Lu, Y. T., Roskin, K. M., Chen, C. F., … Jacob, H. J. (2004). Comparative recombination rates in the rat, mouse, & human genomes. *Genome Research*, *14*, 528–538. https://doi.org/10.1101/gr.1970304

Johnston, S. E., Berenos, C., Slate, J., & Pemberton, J. M. (2016). Conserved genetic architecture underlying individual recombination rate variation in a wild population of soay sheep (*Ovis aries*). *Genetics*, *203*, 583–598. https://doi.org/10.1534/genetics.115.185553

Johnston, S. E., Huisman, J., Ellis, P. A., & Pemberton, J. M. (2017). A high density linkage map reveals sexual dimorphism in recombination landscapes in red deer (*Cervus elaphus*). *G3: Genes, Genomes, Genetics*, *7*, 2859–2870. https://doi.org/10.1534/g3.117.044198

Jones, F. C., Brown, C., Pemberton, J. M., & Braithwaite, V. A. (2006). Reproductive isolation in a threespine stickleback hybrid zone. *Journal of Evolutionary Biology*, *19*, 1531–1544. https://doi.org/10.1111/j.1420-9101.2006.01122.x

Jones, F. C., Chan, Y. F., Schmutz, J., Grimwood, J., Brady, S. D., Southwick, A. M., … Kingsley, D. M. (2012). A genome-wide SNP genotyping array reveals patterns of global and repeated species–pair divergence in sticklebacks. *Current Biology*, *22*, 83–90. https://doi.org/10.1016/j.cub.2011.11.045

Juneja, P., Osei-Poku, J., Ho, Y. S., Ariani, C. V., Palmer, W. J., Pain, A., & Jiggins, F. M. (2014). Assembly of the genome of the disease vector *Aedes aegypti* onto a genetic linkage map allows mapping of genes affecting disease transmission. *PLoS Neglected Tropical Diseases*, *8*, e2652. https://doi.org/10.1371/journal.pntd.0002652

Kaback, D. B., Guacci, V., Barber, D., & Mahon, J. W. (1992). Chromosome size-dependent control of meiotic recombination. *Science*, *256*, 228–232. https://doi.org/10.1126/science.1566070

Kapusta, A., Suh, A., & Feschotte, C. (2017). Dynamics of genome size evolution in birds and mammals. *Proceedings of the National Academy of Sciences of the United States of America*, *114*, E1460–E1469. https://doi.org/10.1073/pnas.1616702114

Kawakami, T., Smeds, L., Backstrom, N., Husby, A., Qvarnstrom, A., Mugal, C. F., … Ellegren, H. (2014). A high-density linkage map enables a second-generation collared flycatcher genome assembly and reveals the patterns of avian recombination rate variation and chromosomal evolution. *Molecular Ecology*, *23*, 4035–4058. https://doi.org/10.1111/mec.12810

Kirkpatrick, M., & Barton, N. (2006). Chromosome inversions, local adaptation and speciation. *Genetics*, *173*, 419–434. https://doi.org/10.1534/genetics.105.047985

Klutstein, M., & Cooper, J. P. (2014). The chromosomal courtship dance – homolog pairing in early meiosis. *Current Opinion in Cell Biology*, *26*, 123–131. https://doi.org/10.1016/j.ceb.2013.12.004

Kochakpour, N., & Moens, P. B. (2008). Sex-specific crossover patterns in Zebrafish (*Danio rerio*). *Heredity*, *100*, 489–495. https://doi.org/10.1038/sj.hdy.6801091

Kondrashov, A. S. (1982). Selection against harmful mutations in large sexual and asexual populations. *Genetics Research*, *40*, 325–332. https://doi.org/10.1017/S0016672300019194

Lambie, E. J., & Roeder, G. S. (1986). Repression of meiotic crossing over by a centromere (CEN3) in *Saccharomyces cerevisiae*. *Genetics*, *114*, 769–789.

Laurent, B., Palaiokostas, C., Spataro, C., Moinard, M., Zehraoui, E., Houston, R. D., & Foulongne-Oriol, M. (2017). High-resolution mapping of

the recombination landscape of the phytopathogen *Fusarium graminearum* suggests two-speed genome evolution. *Molecular Plant Pathology*, https://doi.org/10.1111/mpp.12524

Lee, C. Y., Conrad, M. N., & Dresser, M. E. (2012). Meiotic chromosome pairing is promoted by telomere-led chromosome movements independent of bouquet formation. *PLoS Genetics*, *8*(5), e1002730. https://doi.org/10.1371/journal.pgen.1002730

Lefrancois, P., Rockmill, B., Xie, P. X., Roeder, G. S., & Snyder, M. (2016). Multiple pairwise analysis of non-homologous centromere coupling reveals preferential chromosome size-dependent interactions and a role for bouquet formation in establishing the interaction pattern. *PLoS Genetics*, *12*(10), e1006347. https://doi.org/10.1371/journal.pgen.1006347

Lescak, E. A., Bassham, S. L., Catchen, J., Gelmond, O., Sherbick, M. L., von Hippel, F. A., & Cresko, W. A. (2015). Evolution of stickleback in 50 years on earthquake-uplifted islands. *Proceedings of the National Academy of Sciences of the United States of America*, *112*, E7204–E7212. https://doi.org/10.1073/pnas.1512020112

Levan, A., Fredga, K., & Sandberg, A. A. (1964). Nomenclature for centromeric position on chromosomes. *Hereditas*, *52*, 201–220.

Li, G., Hillier, L. W., Grahn, R. A., Zimin, A. V., David, V. A., Menotti-Raymond, M., … William, J. M. (2016). A high-resolution SNP array-based linkage map anchors a new domestic cat draft genome assembly and provides detailed patterns of recombination. *G3: Genes, Genomes, Genetics*, *6*, 1607–1616. https://doi.org/10.1534/g3.116.028746

Lichten, M., & Goldman, A. S. H. (1995). Meiotic recombination hotspots. *Annual Review of Genetics*, *29*, 423–444. https://doi.org/10.1146/annurev.ge.29.120195.002231

Lien, S., Gidskehaug, L., Moen, T., Hayes, B. J., Berg, P. R., Davidson, W. S., … Kent, M. P. (2011). A dense SNP-based linkage map of Atlantic salmon (*Salmo salar*) reveals extended chromosome homeologies and striking differences in sex-specific recombination patterns. *BMC Genomics*, *12*, 615. https://doi.org/10.1186/1471-2164-12-615

Liu, H., Jia, Y., Sun, X., Tian, D., Hurst, L. D., & Yang, S. (2017). Direct determination of the mutation rate in the bumblebee reveals evidence for weak recombination-associated mutation and an approximate rate constancy in insects. *Molecular Biology and Evolution*, *34*, 119–130. https://doi.org/10.1093/molbev/msw226

Liu, Z., Liu, S., Yao, J., Bao, L., Zhang, J., Li, Y., … Waldbieser, G. C. (2016). The channel catfish genome sequence provides insights into the evolution of scale formation in teleosts. *Nature Communications*, *7*, 11757. https://doi.org/10.1038/ncomms11757

Liu, H., Zhang, X., Huang, J., Chen, J. Q., Tian, D., Hurst, L. D., & Yang, S. (2015). Causes and consequences of crossing-over evidenced via a high-resolution recombinational landscape of the honey bee. *Genome Biology*, *16*, 15. https://doi.org/10.1186/s13059-014-0566-0

Lukaszewski, A. J. (1997). The development and meiotic behavior of asymmetrical isochromosomes in wheat. *Genetics*, *145*, 1155–1160.

Luo, M. C., You, F. M., Li, P., Wang, J. R., Zhu, T., Dandekar, A. M., … Dvorak, J. (2015). Synteny analysis in Rosids with a walnut physical map reveals slow genome evolution in long-lived woody perennials. *BMC Genomics*, *16*, 707. https://doi.org/10.1186/s12864-015-1906-5

Lynch, M., Blanchard, J., Houle, D., Kibota, T., Schultz, S., Vassilieva, L., & Willis, J. (1999). Perspective: Spontaneous deleterious mutation. *Evolution*, *53*, 645–663. https://doi.org/10.1111/j.1558-5646.1999.tb05361.x

Ma, L., O'Connell, J. R., VanRaden, P. M., Shen, B., Padhi, A., Sun, C., … Wiggans, G. R. (2015). Cattle sex-specific recombination and genetic control from a large pedigree analysis. *PLoS Genetics*, *11*, e1005387. https://doi.org/10.1371/journal.pgen.1005387

Mahtani, M. M., & Willard, H. F. (1998). Physical and genetic mapping of the human X chromosome centromere: Repression of recombination. *Genome Research*, *8*, 100–110. https://doi.org/10.1101/gr.8.2.100

Malik, H. S., & Henikoff, S. (2009). Major evolutionary transitions in centromere complexity. *Cell*, *138*, 1067–1082. https://doi.org/10.1016/j.cell.2009.08.036

Manly, B. F. J. (2007). *Randomization, bootstrap and Monte Carlo methods in biology*, 3rd edn. Boca Raton, FL: Chapman & Hall.

Mather, K. (1938). Crossing-over. *Biological Reviews*, *13*, 252–292. https://doi.org/10.1111/j.1469-185X.1938.tb00516.x

Maynard Smith, J., & Haigh, J. (1974). Hitch-hiking effect of a favorable gene. *Genetics Research*, *23*, 23–35. https://doi.org/10.1017/S0016672300014634

McFarlane, R. J., & Humphrey, T. C. (2010). A role for recombination in centromere function. *Trends in Genetics*, *26*, 209–213. https://doi.org/10.1016/j.tig.2010.02.005

Melters, D. P., Paliulis, L. V., Korf, I. F., & Chan, S. W. L. (2012). Holocentric chromosomes: Convergent evolution, meiotic adaptations, & genomic analysis. *Chromosome Research*, *20*, 579–593. https://doi.org/10.1007/s10577-012-9292-1

Mezard, C. (2006). Meiotic recombination hotspots in plants. *Biochemical Society Transactions*, *34*, 531–534. https://doi.org/10.1042/BST0340531

Muller, H. J. (1916). The mechanism of crossing-over. *American Naturalist*, *50*, 193–221. https://doi.org/10.1086/279534

Muller, H. J. (1932). Some genetic aspects of sex. *American Naturalist*, *66*, 118–138. https://doi.org/10.1086/280418

Muñoz-Fuentes, V., Marcet-Ortega, M., Alkorta-Aranburu, G., Linde Forsberg, C., Morrell, J. M., Manzano-Piedras, E., … Vila, C. (2015). Strong artificial selection in domestic mammals did not result in an increased recombination rate. *Molecular Biology and Evolution*, *32*, 510–523. https://doi.org/10.1093/molbev/msu322

Nachman, M. W., & Churchill, G. A. (1996). Heterogeneity in rates of recombination across the mouse genome. *Genetics*, *142*, 537–548.

Nachman, M. W., & Payseur, B. A. (2012). Recombination rate variation and speciation: Theoretical predictions and empirical results from rabbits and mice. *Philosophical Transactions of the Royal Society B*, *367*, 409–421. https://doi.org/10.1098/rstb.2011.0249

Nam, K., & Ellegren, H. (2012). Recombination drives vertebrate genome contraction. *PLoS Genetics*, *8*(5), e1002680. https://doi.org/10.1371/journal.pgen.1002680

Naranjo, T., & Corredor, E. (2008). Nuclear architecture and chromosome dynamics in the search of the pairing partner in meiosis in plants. *Cytogenetic and Genome Research*, *120*, 320–330. https://doi.org/10.1159/000121081

Niehuis, O., Gibson, J. D., Rosenberg, M. S., Pannebakker, B. A., Koevoets, T., Judson, A. K., … Gadau, J. (2010). Recombination and its impact on the genome of the haplodiploid parasitoid wasp *Nasonia*. *PLoS ONE*, *5*, e8597. https://doi.org/10.1371/journal.pone.0008597

Noor, M. A. F., & Bennett, S. M. (2009). Islands of speciation or mirages in the desert? Examining the role of restricted recombination in maintaining species. *Heredity*, *103*, 439–444.

Nordborg, M., Charlesworth, B., & Charlesworth, D. (1996). The effect of recombination on background selection. *Genetics Research*, *67*, 159–174. https://doi.org/10.1017/S0016672300033619

Nunes, J. D. D., Liu, S. K., Pertille, F., Perazza, C. A., Villela, P. M. S., de Almeida-Val, V. M. F., … Coutinho, L. L. (2017). Large-scale SNP discovery and construction of a high-density genetic map of *Colossoma macropomum* through genotyping-by-sequencing. *Scientific Reports*, *7*, 46112. https://doi.org/10.1038/srep46112

Ortiz-Barrientos, D., Reiland, J., Hey, J., & Noor, M. A. F. (2002). Recombination and the divergence of hybridizing species. *Genetica*, *116*, 167–178. https://doi.org/10.1023/A:1021296829109

Otto, S. P., & Barton, N. H. (1997). The evolution of recombination: Removing the limits to natural selection. *Genetics*, *147*, 879–906.

Otto, S. P., & Barton, N. H. (2001). Selection for recombination in small populations. *Evolution*, *55*, 1921–1931. https://doi.org/10.1111/j.0014-3820.2001.tb01310.x

Page, S. L., & Hawley, R. S. (2003). Chromosome choreography: The meiotic ballet. *Science*, *301*, 785–789. https://doi.org/10.1126/science.1086605

Payseur, B. A., & Nachman, M. W. (2002). Gene density and human nucleotide polymorphism. *Molecular Biology and Evolution*, *19*, 336–340. https://doi.org/10.1093/oxfordjournals.molbev.a004086

Perreault-Payette, A., Muir, A. M., Goetz, F., Perrier, C., Normandeau, E., Sirois, P., & Bernatchez, L. (2017). Investigating the extent of parallelism in morphological and genomic divergence among lake trout ecotypes in Lake Superior. *Molecular Ecology*, *26*, 1477–1497. https://doi.org/10.1111/mec.14018

Petkov, P. M., Broman, K. W., Szatkiewicz, J. P., & Paigen, K. (2007). Crossover interference underlies sex differences in recombination rates. *Trends in Genetics*, *23*, 539–542. https://doi.org/10.1016/j.tig.2007.08.015

Pratto, F., Brick, K., Khil, P., Smagulova, F., Petukhova, G. V., & Camerini-Otero, R. D. (2014). Recombination initiation maps of individual human genomes. *Science*, *346*, 1256442. https://doi.org/10.1126/science.1256442

Puchta, H. (2005). The repair of double-strand breaks in plants: Mechanisms and consequences for genome evolution. *Journal of Experimental Botany*, *56*, 1–14.

R Core Team (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Raeymaekers, J. A. M., Chaturvedi, A., Hablützel, P. I., Verdonck, I., Hellemans, B., Maes, G. E., … Volckaert, F. A. M. (2017). Adaptive and non–adaptive divergence in a common landscape. *Nature Communications*, *8*, 267. https://doi.org/10.1038/s41467-017-00256-6

Rahn, M. I., & Solari, A. J. (1986). Recombination nodules in the oocytes of the chicken, *Gallus domesticus*. *Cytogenetic and Cell Genetics*, *43*, 187–193. https://doi.org/10.1159/000132319

Ravinet, M., Westram, A., Johannesson, K., Butlin, R., Andre, C., & Panova, M. (2016). Shared and nonshared genomic divergence in parallel ecotypes of *Littorina saxatilis* at a local scale. *Molecular Ecology*, *25*, 287–305. https://doi.org/10.1111/mec.13332

Rees, H., & Dale, P. J. (1974). Chiasmata and variability in *Lolium* and *Festuca* populations. *Chromosoma*, *47*, 335–351. https://doi.org/10.1007/BF00328866

Ren, Y., Zhao, H., Kou, Q., Jiang, J., Guo, S., Zhang, H., … Xu, Y. (2012). A high resolution genetic map anchoring scaffolds of the sequenced watermelon genome. *PLoS ONE*, *7*, e29453. https://doi.org/10.1371/journal.pone.0029453

Renaut, S., Grassa, C. J., Yeaman, S., Moyers, B. T., Lai, Z., Kane, N. C., … Rieseberg, L. H. (2013). Genomic islands of divergence are not affected by geography of speciation in sunflowers. *Nature Communications*, *4*, 1827. https://doi.org/10.1038/ncomms2833

Renaut, S., Owens, G. L., & Rieseberg, L. H. (2014). Shared selective pressure and local genomic landscape lead to repeatable patterns of genomic divergence in sunflowers. *Molecular Ecology*, *23*, 311–324. https://doi.org/10.1111/mec.12600

Rockman, M. V., & Kruglyak, L. (2009). Recombinational landscape and population genomics of *Caenorhabditis elegans*. *PLoS Genetics*, *5*, e1000419.

Roesti, M., Gavrilets, S., Hendry, A. P., Salzburger, W., & Berner, D. (2014). The genomic signature of parallel adaptation from shared genetic variation. *Molecular Ecology*, *23*, 3944–3956. https://doi.org/10.1111/mec.12720

Roesti, M., Hendry, A. P., Salzburger, W., & Berner, D. (2012). Genome divergence during evolutionary diversification as revealed in replicate lake-stream stickleback population pairs. *Molecular Ecology*, *21*, 2852–2862. https://doi.org/10.1111/j.1365-294X.2012.05509.x

Roesti, M., Kueng, B., Moser, D., & Berner, D. (2015). The genomics of ecological vicariance in threespine stickleback fish. *Nature Communications*, *6*, 8767. https://doi.org/10.1038/ncomms9767

Roesti, M., Moser, D., & Berner, D. (2013). Recombination in the threespine stickleback genome – patterns and consequences. *Molecular Ecology*, *22*, 3014–3027. https://doi.org/10.1111/mec.12322

Ross, J. A., Koboldt, D. C., Staisch, J. E., Chamberlin, H. M., Gupta, B. P., Miller, R. D., … Haag, E. S. (2011). Caenorhabditis briggsae recombinant inbred line genotypes reveal inter-strain incompatibility and the evolution of recombination. *PLoS Genetics*, *7*, e1002174. https://doi.org/10.1371/journal.pgen.1002174

Ross-Ibarra, J. (2004). The evolution of recombination under domestication: A test of two hypotheses. *American Naturalist*, *163*, 105–112. https://doi.org/10.1086/380606

Rougemont, Q., Gagnaire, P. A., Perrier, C., Genthon, C., Besnard, A. L., Launey, S., & Evanno, G. (2017). Inferring the demographic history underlying parallel genomic divergence among pairs of parasitic and nonparasitic lamprey ecotypes. *Molecular Ecology*, *26*, 142–162. https://doi.org/10.1111/mec.13664

Samuk, K., Owens, G. L., Delmore, K. E., Miller, S. E., Rennison, D. J., & Schluter, D. (2017). Gene flow and selection interact to promote adaptive divergence in regions of low recombination. *Molecular Ecology*, *17*, 4378–4390. https://doi.org/10.1111/mec.14226

Scherthan, H. (2001). A bouquet makes ends meet. *Nature Reviews Molecular Cell Biology*, *2*, 621–627. https://doi.org/10.1038/35085086

Scherthan, H., Weich, S., Schwegler, H., Heyting, C., Harle, M., & Cremer, T. (1996). Centromere and telomere movements during early meiotic prophase of mouse and man are associated with the onset of chromosome pairing. *Journal of Cell Biology*, *134*, 1109–1125. https://doi.org/10.1083/jcb.134.5.1109

Schmutz, J., Cannon, S. B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., … Jackson, S. A. (2010). Genome sequence of the palaeopolyploid soybean. *Nature*, *463*, 178–183. https://doi.org/10.1038/nature08670

Schnable, P. S., Hsia, A. P., & Nikolau, B. J. (1998). Genetic recombination in plants. *Current Opinions in Plant Biology*, *1*, 123–129. https://doi.org/10.1016/S1369-5266(98)80013-7

Schubert, I., & Vu, G. T. H. (2016). Genome stability and evolution: Attempting a holistic view. *Trends Plant Sciences*, *21*, 749–757. https://doi.org/10.1016/j.tplants.2016.06.003

Sherman, J. D., & Stack, S. M. (1995). Two–dimensional spreads of synaptonemal complexes from solanaceous plants. VI. High-resolution recombination nodule map for tomato (*Lycopersicon esculentum*). *Genetics*, *141*, 683–708.

Shriver, M. D., Smith, M. W., Jin, L., Marcini, A., Akey, J. M., Deka, R., & Ferrell, R. E. (1997). Ethnic-affiliation estimation by use of population-specific DNA markers. *American Journal of Human Genetics*, *60*, 957–964.

Shulaev, V., Sargent, D. J., Crowhurst, R. N., Mockler, T. C., Folkerts, O., Delcher, A. L., … Folta, K. M. (2011). The genome of woodland strawberry (*Fragaria vesca*). *Nature Genetics*, *43*, 109–116. https://doi.org/10.1038/ng.740

Sim, S. B., & Geib, S. M. (2017). A chromosome-scale assembly of the *Bactrocera cucurbitae* genome provides insight to the genetic basis of white pupae. *G3: Genes, Genomes, Genetics*, *7*, 1927–1940.

Sirviö, A., Gadau, J., Rueppell, O., Lamatsch, D., Boomsma, J. J., Pamilo, P., & Page, R. E. (2006). High recombination frequency creates genotypic diversity in colonies of the leaf-cutting ant *Acromyrmex echinatior*. *Journal of Evolutionary Biology*, *19*, 1475–1485. https://doi.org/10.1111/j.1420-9101.2006.01131.x

Sirviö, A., Johnston, J. S., Wenseleers, T., & Pamilo, P. (2011b). A high recombination rate in eusocial Hymenoptera: Evidence from the common wasp *Vespula vulgaris*. *BMC Genetics*, *12*, 95. https://doi.org/10.1186/1471-2156-12-95

Sirviö, A., Pamilo, P., Johnson, R. A., Page, R. E. Jr, & Gadau, J. (2011a). Origin and evolution of the dependent lineages in the genetic caste determination system of *Pogonomyrmex* ants. *Evolution*, *65*, 869–884. https://doi.org/10.1111/j.1558-5646.2010.01170.x

Smeds, L., Mugal, C. F., Qvarnstrom, A., & Ellegren, H. (2016). High-resolution mapping of crossover and non-crossover recombination events by whole-genome re-sequencing of an avian pedigree. *PLoS Genetics*, *12*, e1006044. https://doi.org/10.1371/journal.pgen.1006044

Smith, K. N., & Nicolas, A. (1998). Recombination at work for meiosis. *Current Opinion in Genetics & Development*, *8*, 200–211. https://doi.org/10.1016/S0959-437X(98)80142-1

Solignac, M., Mougel, F., Vautrin, D., Monnerot, M., & Cornuet, J. M. (2007). A third-generation microsatellite–based linkage map of the honey bee, *Apis mellifera*, and its comparison with the sequence-based physical map. *Genome Biology*, *8*, R66. https://doi.org/10.1186/gb-2007-8-4-r66

Stapley, J., Feulner, P. G. D., Johnston, S. E., Santure, A. W., & Smadja, C. M. (2017). Variation in recombination frequency and distribution across eukaryotes: Patterns and processes. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*, 20160455. https://doi.org/10.1098/rstb.2016.0455

Stuart, Y. E., Veen, T., Weber, J. N., Hanson, D., Ravinet, M., Lohman, B. K., … Bolnick, D. I. (2017). Contrasting effects of environment and genetics generate a continuum of parallel evolution. *Nature Ecology and Evolution*, *1*, 0158. https://doi.org/10.1038/s41559-017-0158

Sturtevant, A. H. (1915). The behavior of the chromosomes as studied through linkage. *Zeitschrift für indukive Abstammungs- und Vererbungslehre*, *13*, 234–287.

Talbert, P. B., & Henikoff, S. (2010). Centromeres convert but don't cross. *PLoS Biology*, *8*, e1000326.

Tease, C., & Hulten, M. A. (2004). Inter-sex variation in synaptonemal complex lengths largely determine the different recombination rates in male and female germ cells. *Cytogenetic and Genome Research*, *107*, 208–215. https://doi.org/10.1159/000080599

Terekhanova, N. V., Logacheva, M. D., Penin, A. A., Neretina, T. V., Barmintseva, A. E., Bazykin, G. A., … Mugue, N. S. (2014). Fast evolution from precast bricks: Genomics of young freshwater populations of threespine stickleback *Gasterosteus aculeatus*. *PLoS Genetics*, *10*, e1004696. https://doi.org/10.1371/journal.pgen.1004696

Tian, Z., Rizzon, C., Du, J., Zhu, L., Bennetzen, J. L., Jackson, S. A., … Ma, J. (2009). Do genetic recombination and gene density shape the pattern of DNA elimination in rice long terminal repeat retrotransposons? *Genome Research*, *19*, 2221–2230. https://doi.org/10.1101/gr.083899.108

Tine, M., Kuhl, H., Gagnaire, P.-N. D. A. S. H.-A., Louro, B., Desmarais, E., Martins, R. S. T., … Reinhardt, R. (2014). European sea bass genome and its variation provide insights into adaptation to euryhalinity and speciation. *Nature Communications*, *5*, 5770. https://doi.org/10.1038/ncomms6770

Tomato Genome Consortium (2012). The tomato genome sequence provides insights into fleshy fruit evolution. *Nature*, *485*, 635–641.

Tong, C., Li, H., Wang, Y., Li, X., Ou, J., Wang, D., … Shi, J. (2016). Construction of high-density linkage maps of *Populus deltoides* × *P. simonii* using restriction-site associated DNA sequencing. *PLoS ONE*, *11*, e0150692. https://doi.org/10.1371/journal.pone.0150692

Tortereau, F., Servin, B., Frantz, L., Megens, H.-J., Milan, D., Rohrer, G., … Groenen, M. (2012). A high density recombination map of the pig reveals a correlation between sex-specific recombination and GC content. *BMC Genomics*, *13*, 586. https://doi.org/10.1186/1471-2164-13-586

Trucchi, E., Frajman, B., Haverkamp, T. H. A., Schönswetter, P., & Paun, O. (2017). Genomic analyses suggest parallel ecological divergence in *Heliosperma pusillum* (Caryophyllaceae). *New Phytologist*, *216*, 267–278. https://doi.org/10.1111/nph.14722

Van Doren, B. M., Campagna, L., Helm, B., Illera, J. C., Lovette, I. J., & Liedvogel, M. (2017). Correlated patterns of genetic diversity and differentiation across an avian family. *Molecular Ecology*, *26*, 3982–3997. https://doi.org/10.1111/mec.14083

Viera, A., Santos, J. L., & Rufas, J. S. (2009). Relationship between incomplete synapsis and chiasma localization. *Chromosoma*, *118*, 377–389. https://doi.org/10.1007/s00412-009-0204-x

Vincenten, N., Kuhl, L. M., Lam, I., Oke, A., Kerr, A. R. W., Hochwagen, A., … Marston, A. L. (2015). The kinetochore prevents centromere–proximal crossover recombination during meiosis. *eLife*, *4*, e10850.

Wang, L., Bai, B., Liu, P., Huang, S. Q., Wan, Z. Y., Chua, E., … Yue, G. H. (2017). Construction of high-resolution recombination maps in Asian seabass. *BMC Genomics*, *18*, 63. https://doi.org/10.1186/s12864-016-3462-z

Wang, S., Chen, J., Zhang, W., Hu, Y., Chang, L., Fang, L., … Thang, T. (2015a). Sequence-based ultra-dense genetic and physical maps reveal structural variations of allopolyploid cotton genomes. *Genome Biology*, *16*, 108. https://doi.org/10.1186/s13059-015-0678-1

Wang, X., Yu, K., Li, H., Peng, Q., Chen, F., Zhang, W., … Zhang, J. (2015b). High-density SNP map construction and QTL identification for the apetalous character in *Brassica napus* L. *Frontiers in Plant Science*, *6*, 1164.

Wilfert, L., Gadau, J., & Schmid-Hempel, P. (2007). Variation in genomic recombination rates among animal taxa and the case of social insects. *Heredity*, *98*, 189–197. https://doi.org/10.1038/sj.hdy.6800950

Wolf, K. W. (1994). How meiotic cells deal with non-exchange chromosomes. *BioEssays*, *16*, 107–114. https://doi.org/10.1002/(ISSN)1521-1878

Wong, A. K., Ruhe, A. L., Dumont, B. L., Robertson, K. R., Guerrero, G., Shull, S. M., … Neff, M. W. (2010). A comprehensive linkage map of the dog genome. *Genetics*, *184*, 595–U436. https://doi.org/10.1534/genetics.109.106831

Wu, G. A., Prochnik, S., Jenkins, J., Salse, J., Hellsten, U., Murat, F., … Rokhsar, D. (2014). Sequencing of diverse mandarin, pummelo and orange genomes reveals complex history of admixture during citrus domestication. *Nature Biotechnology*, *32*, 656–662. https://doi.org/10.1038/nbt.2906

Xiang, Y. B., Miller, D. E., Ross, E. J., Alvarado, A. S., & Hawley, R. S. (2014). Synaptonemal complex extension from clustered telomeres mediates full-length chromosome pairing in *Schmidtea mediterranea*. *Proceedings of the National Academy of Sciences of the United States of America*, *111*, E5159–E5168. https://doi.org/10.1073/pnas.1420287111

Xu, P., Zhang, X., Wang, X., Li, J., Liu, G., Kuang, Y., … Sun, X. (2014). Genome sequence and genetic diversity of the common carp, *Cyprinus carpio*. *Nature Genetics*, *46*, 1212–1219. https://doi.org/10.1038/ng.3098

Yan, H. H., Jin, W. W., Nagaki, K., Tian, S. L., Ouyang, S., Buell, C. R., … Jiang, J. M. (2005). Transcription and histone modifications in the recombination–free region spanning a rice centromere. *Plant Cell*, *17*, 3227–3238. https://doi.org/10.1105/tpc.105.037945

Zhang, G., Liu, X., Quan, Z., Cheng, S., Xu, X., Pan, S., … Wang, J. (2012). Genome sequence of foxtail millet (*Setaria italica*) provides insights into grass evolution and biofuel potential. *Nature Biotechnology*, *30*, 549–554. https://doi.org/10.1038/nbt.2195

Zickler, D., & Kleckner, N. (2016). A few of our favorite things: Pairing, the bouquet, crossover interference and evolution of meiosis. *Seminars in Cell and Developmental Biology*, *54*, 135–148. https://doi.org/10.1016/j.semcdb.2016.02.024

SUPPORTING INFORMATION

Additional supplemental material may be found online in the Supporting Information section at the end of the article.

How to cite this article: Haenel Q, Laurentino TG, Roesti M, Berner D. Meta-analysis of chromosome-scale crossover rate variation in eukaryotes and its significance to evolutionary genomics. *Mol Ecol*. 2018;27:2477–2497. https://doi.org/10.1111/mec.14699

MOLECULAR ECOLOGY

Appendix 1 to:

Meta-analysis of chromosome-scale crossover rate variation in eukaryotes and its significance to evolutionary genomics

by Quiterie HAENEL, Telma G. LAURENTINO, Marius ROESTI & Daniel BERNER

Contents:

Table S1.

References used to obtain information on centromere position in the organisms indicated in Table 1 in the main paper. The references are sorted alphabetically by species name.

Aedes aegypti
Sharakhova, M. V. et al. (2011) Imaginal discs - A new source of chromosomes for genome mapping of the yellow fever mosquito *Aedes aegypti*. PLoS Negl. Trop. Dis. 5, 1–9.

Beta vulgaris
Paesold, S., Borchardt, D., Schmidt, T. & Dechyeva, D. (2012) A sugar beet (*Beta vulgaris* L.) reference FISH karyotype for chromosome and chromosome-arm identification, integration of genetic linkage groups and analysis of major repeat family distribution. Plant J. 72, 600–611.

Bos taurus
Di Berardino, D., Di Meo, G. P., Gallagher, D. S., Hayes, H. & Iannuzzi, L. (co-ordinator) (2001). ISCNDB 2000 International System for Chromosome Nomenclature of Domestic Bovids. Cytogenet. Cell Genet. 299, 283–299.

Caenorhabditis briggsae, C. elegans
Hillier, L. D. W. et al. (2007) Comparison of *C. elegans* and *C. briggsae* genome sequences reveals extensive conservation of chromosome organization and synteny. PLoS Biol. 5, 1603–1616.

Canis lupus familiaris
Yang, F. et al. (2000) Chromosome identification and assignment of DNA clones in the dog using a red fox and dog comparative map. Chromosom. Res. 8, 93–100.

Cervus elaphus
Johnston, S. E., Huisman, J., Ellis, P. A. & Pemberton, J. M. (2017) A high-density linkage map reveals sexually-dimorphic recombination landscapes in red deer (*Cervus elaphus*). *G3: Genes*|*Genomes*|*Genetics* 7, 2859–2870.

Felis catus
Yang, F. et al. (2000) Reciprocal chromosome painting illuminates the history of genome evolution of the domestic cat, dog and human. Chromosom. Res. 8, 393–404.

Gasterosteus aculeatus
Urton JR, McCann SM, Peichel CL (2011) Karyotype differentiation between two stickleback species (Gasterosteidae). Cytogenet. Genome Res. 135, 150-159.

Heliconius melpomene
Ahola, V. et al. (2014) The Glanville fritillary genome retains an ancient karyotype and reveals selective chromosomal fusions in Lepidoptera. Nat. Commun. 5, 4737.

Homo sapiens
http://www.ensembl.org/Homo_sapiens/Location/Genome

Mus musculus
http://www.ensembl.org/Mus_musculus/Location/Genome

Nasonia vitripennis
Rutten, K. B. et al. (2004) Chromosomal anchoring of linkage groups and identification of wing size QTL using markers and FISH probes derived from microdissected chromosomes in *Nasonia* (Pteromalidae: Hymenoptera). Cytogenet. Genome Res. 105, 126–133.

Ovis aries
Di Berardino, D., Di Meo, G. P., Gallagher, D. S., Hayes, H. & Iannuzzi, L. (co-ordinator) (2001). ISCNDB 2000 International System for Chromosome Nomenclature of Domestic Bovids. Cytogenet. Cell Genet. 299, 283–299.
Goldammer, T. et al. (2009) Molecular cytogenetics and gene mapping in sheep (*Ovis aries*, 2n = 54). Cytogenet. Genome Res. 126, 63–76.

Pan troglodytes verus
Lin, C. C., Chiarelli, B., Cohen, M. & Boer, L. E. M. de. (1973) A comparison of the fluorescent karyotypes of the chimpanzee (*Pan troglodytes*) and man. J. Hum. Evol. 2, 311–321.

Phaseolus vulgaris
Fonsêca, A. et al. (2010) Cytogenetic map of common bean (*Phaseolus vulgaris* L.). Chromosom. Res. 18, 487–502.

Rattus norvegicus
Hamta, A. et al. (2006) Chromosome ideograms of the laboratory rat (*Rattus norvegicus*) based on high-resolution banding, and anchoring of the cytogenetic map to the DNA sequence by FISH in sample chromosomes. Cytogenet. Genome Res. 115, 158–168.

Figure S1

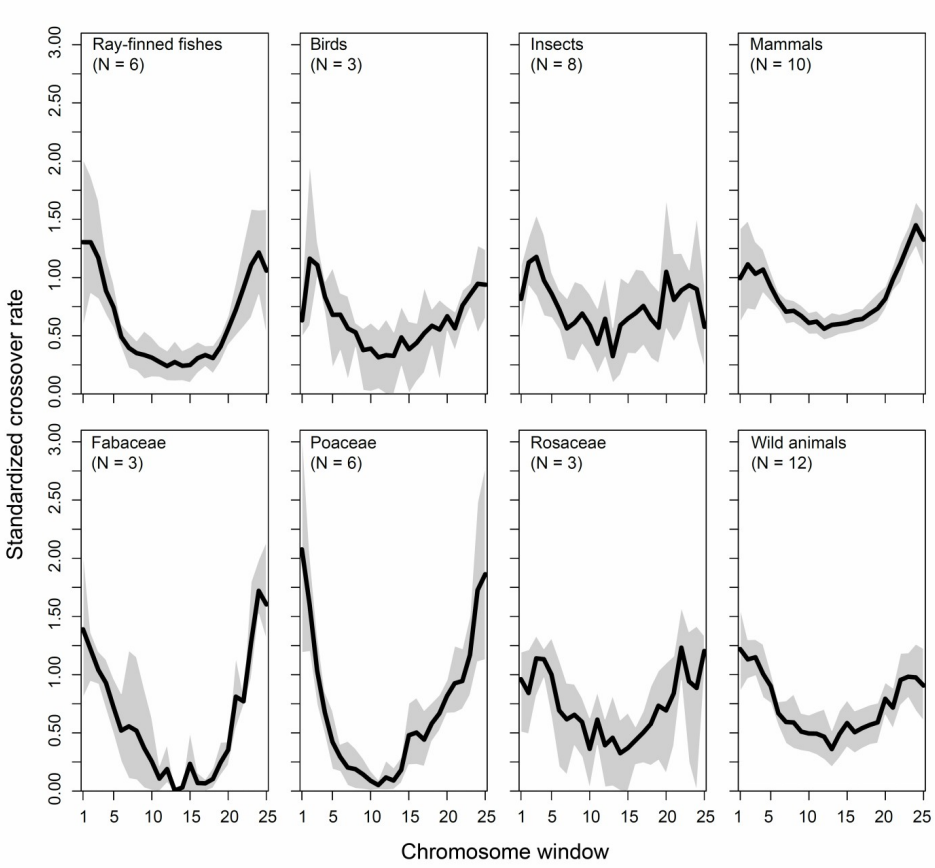


Fig. S1. A reduced crossover (CO) rate around chromosome centers is well supported across taxonomic groups within our focal eukaryote kingdoms, as shown by average CO rate profiles generated separately for selected animal classes (top row) and plant families (bottom row, first three panels from the left). A similar CO distribution also emerges when restricting the analysis to wild species only (i.e., domesticated species excluded), shown in the bottom right panel (for animals only; wild plant species were too few in our data base). Sample sizes (i.e., number of species) are given in parentheses. All plotting conventions follow Figure 1 in the main paper.

Figure S2

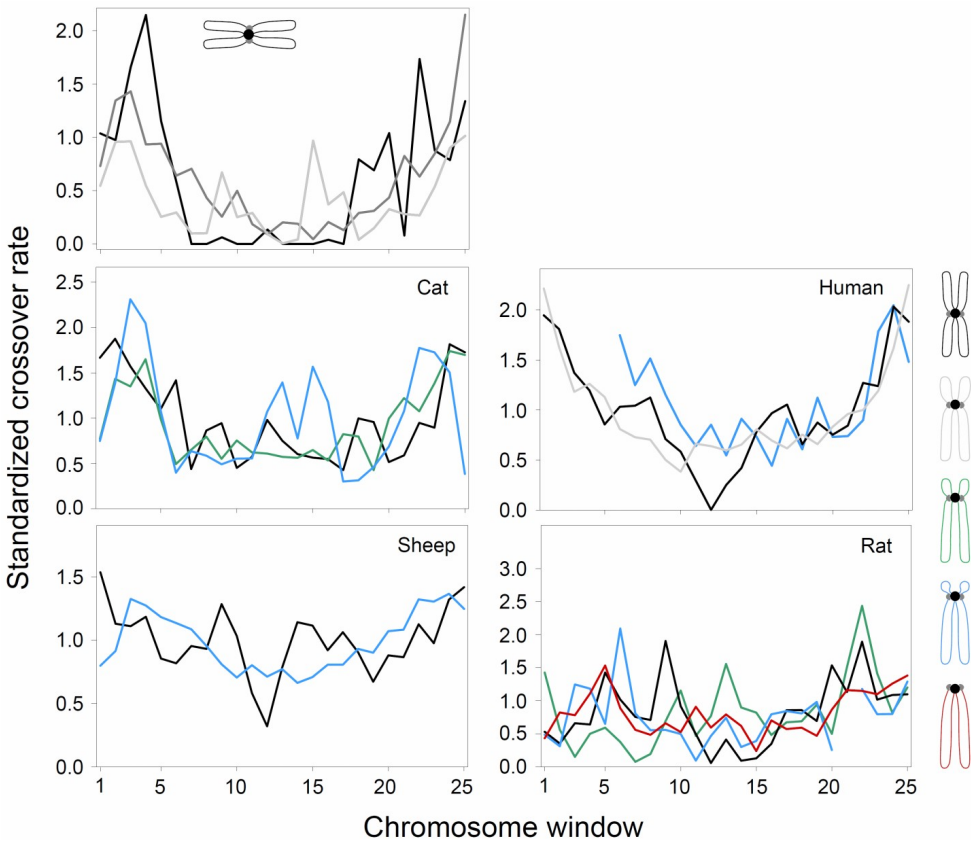


Fig. S2. The top left panel shows the distribution of CO rate in three species with metacentric chromosomes (*Nasonia* wasp, Sugar beet, and Yellow fever mosquito in ligth gray, dark gray and black), the other panels show CO rate profiles separately for different chromosome morphologies within four species. Color coding and plotting conventions follow Figure 2 in the main paper.

Figure S3

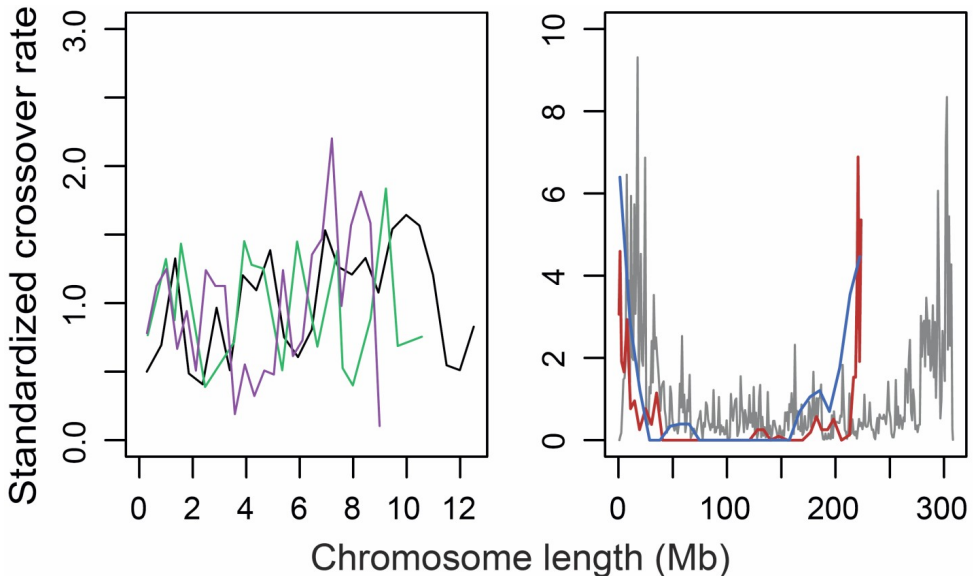


Fig. S3. Short chromosomes generally display a relatively uniform CO distribution, as illustrated by a single representative chromosome from three species with short chromosomes (left panel; black profile: Honeybee, chromosome 5; green: Monkey flower, chromosome 11; purple: Postman butterfly, chromosome 3). By contrast, long chromosomes typically cross over primarily toward their tips and thus exhibit a vast region of very low CO around their center, exemplified in three species with long chromosomes (right panel; gray: Pig, chromosome 1; red: Pepper, chromosome 5; blue: Maize, chromosome 5). The profiles show mean-standardized CO rates for marker intervals along the chromosome at the original physical scale. Note that the scale of the X-axis is more than 25 times larger in the right than the left panel! The references to the specific studies are provided in Table 1.

4 Outreach

Chapter 5

Biodiversity and community structure of Meiofaunal eukaryotes in shell sand from Hållö island, Smögen, and soft mud from Gullmarn Fjord, Lysekil

Haenel et al. 2017, Biodiversity Data Journal

Biodiversity Data Journal 5: e12731
doi: 10.3897/BDJ.5.e12731

Research Article

NGS-based biodiversity and community structure analysis of meiofaunal eukaryotes in shell sand from Hållö island, Smögen, and soft mud from Gullmarn Fjord, Sweden

Quiterie Haenel‡, Oleksandr Holovachov§, Ulf Jondelius§, Per Sundberg|,¶, Sarah J. Bourlat|,¶

‡ Zoological Institute, University of Basel, Basel, Switzerland
§ Swedish Museum of Natural History, Stockholm, Sweden
| Department of Marine Sciences, University of Gothenburg, Gothenburg, Sweden
¶ SeAnalytics AB, Bohus-Björkö, Sweden

Corresponding author: Sarah J. Bourlat (sarah.bourlat@gmail.com)

Academic editor: Urmas Kõljalg

Received: 15 Mar 2017 | Accepted: 06 Jun 2017 | Published: 08 Jun 2017

Citation: Haenel Q, Holovachov O, Jondelius U, Sundberg P, Bourlat S (2017) NGS-based biodiversity and community structure analysis of meiofaunal eukaryotes in shell sand from Hållö island, Smögen, and soft mud from Gullmarn Fjord, Sweden. Biodiversity Data Journal 5: e12731. https://doi.org/10.3897/BDJ.5.e12731

Abstract

Aim: The aim of this study was to assess the biodiversity and community structure of Swedish meiofaunal eukaryotes using metabarcoding. To validate the reliability of the metabarcoding approach, we compare the taxonomic resolution obtained using the mitochondrial cytochrome oxidase 1 (COI) 'mini-barcode' and nuclear 18S small ribosomal subunit (18S) V1-V2 region, with traditional morphology-based identification of Xenacoelomorpha and Nematoda.

Location: 30 samples were analysed from two ecologically distinct locations along the west coast of Sweden. 18 replicate samples of coarse shell sand were collected along the north-eastern side of Hållö island near Smögen, while 12 replicate samples of soft mud were collected in the Gullmarn Fjord near Lysekil.

Methods: Meiofauna was extracted using flotation and siphoning methods. Both COI and 18S regions were amplified from total DNA samples using Metazoan specific primers and subsequently sequenced using Illumina MiSeq, producing in total 24 132 875 paired-end reads of 300 bp in length, of which 15 883 274 COI reads and 8 249 601 18S reads. These were quality filtered resulting in 7 954 017 COI sequences and 890 370 18S sequences, clustered into 2805 and 1472 representative OTUs respectively, yielding 190 metazoan OTUs for COI and 121 metazoan OTUs for 18S using a 97% sequence similarity threshold.

Results: The Metazoan fraction represents 7% of the total dataset for COI (190 OTUs) and 8% of sequences for 18S (121 OTUs). Annelida (30% of COI metazoan OTUs and 23.97% of 18S metazoan OTUs) and Arthropoda (27.37% of COI metazoan OTUs and 11.57% of 18S metazoan OTUs), were the most OTU rich phyla identified in all samples combined. As well as Annelida and Arthropoda, other OTU rich phyla represented in our samples include Mollusca, Platyhelminthes and Nematoda. In total, 213 COI OTUs and 243 18S OTUs were identified to species using a 97% sequence similarity threshold, revealing some non-native species and highlighting the potential of metabarcoding for biological recording. Taxonomic community composition shows as expected clear differentiation between the two habitat types (soft mud versus coarse shell sand), and diversity observed varies according to choice of meiofaunal sampling method and primer pair used.

Keywords

Meiofaunal biodiversity, community structure, Illumina Mi-Seq, Metabarcoding, COI, 18S

Introduction

Microscopic interstitial marine organisms, also termed ‘meiofauna’, are often defined as animals that pass a 1mm mesh but are retained on a 45 µm sieve (Higgins 1988). Meiofauna are an important component of sedimentary and benthic habitats due to their small size, abundance and rapid turnover rates. Moreover, meiofaunal surveys represent a useful tool for environmental impact assessments, underlying the urgent need for reliable, reproducible and rapid analytical methods. The breadth of taxonomic groups present in marine sediments makes meiofauna an ideal tool for detecting the effects of ecological impacts on marine biodiversity (Moreno et al. 2008). However, traditional morphology based taxonomy assignment methods are labour intensive and time consuming, leading us to explore recently developed metabarcoding methods for whole community analysis. Metabarcoding has previously been used to characterize plankton assemblages (Lindeque et al. 2013, de Vargas et al. 2015), marine benthic meiofaunal assemblages (Creer et al. 2010, Fonseca et al. 2014, Fonseca et al. 2010, Brannock and Halanych 2015, Cowart et al. 2015), meiofaunal communities colonizing autonomous reef monitoring structures (Leray and Knowlton 2015) or fish gut contents (Leray et al. 2013). The vast majority of studies have employed Roche 454 due to its long read lengths compared to other technologies (Table 1; Shokralla et al. 2012), but Illumina MiSeq is now able to provide

similarly long reads using paired-end sequencing (2x300 base pairs). As summarized in Table 1, there is no standardized method for metabarcoding of marine fauna, and a variety of sample extraction methods, sequencing platforms, molecular markers, bioinformatics pipelines and OTU clustering thresholds have been used to date, making these studies difficult to compare (Table 1).

Table 1.

Methodological comparison of benthic and pelagic metabarcoding studies of marine fauna published to date

Authors	Sample type	Sample extraction method	Sequencing platform	Marker	Marker size (bp)	Chimera screening	OTU clustering method and threshold	Database
Leray et al. 2013	Coral reef fish gut contents	Dissection of fish gut	Roche 454 GS FLX	COI	313	UCHIME	CROP 92-94%	Moorea Biocode Database, GenBank
Leray and Knowlton 2015	Autonomous reef monitoring structures	4 fractions (Sessile, 2mm, 500µm, 106µm)	Ion Torrent	COI	313			BOLD, GenBank
Lindeque et al. 2013	Zooplankton from 50m to the surface	200µm mesh WP2 plankton net	Roche 454 GS FLX	18S (V1-V2 regions)	450	ChimeraSlayer (QIIME 1.3.0)	UCLUST 97% (QIIME 1.3.0)	Silva 108, GenBank
de Vargas et al. 2015	Plankton	3 fractions (5-20µm, 20-180µm, 180-2000µm)	Paired-end Illumina Genome Analyser IIx system	18S (V9 region)		USEARCH		V9_PR2, V9 rDNA, Protistan Ribosomal Reference Database
Fonseca et al. 2010	Marine benthic meiofauna	Decanting 45µm sieve Ludox	Roche 454 GS FLX	18S (V1-V2 regions)	364 (250-500)	OCTOPUS	OCTOPUS 96%	GenBank
Fonseca et al. 2014	Marine benthic meiofauna	Decanting 45µm sieve Ludox	Roche 454 GS FLX	18S (V1-V2 regions)	450	Amplicon-Noise	Amplicon-Noise 99% and 96%	GenBank
Brannock and Halanych 2015	Marine benthic meiofauna	Directly from sediment, elutriated on 45µm sieve	Paired-end 100 bp reads Illumina HiSeq	18S (V9 region)	87-187 [13]	USEARCH 6.1. (QIIME 1.8)	UPARSE 97% UCLUST and USEARCH (QIIME 1.8)	Silva 111

Cowart et al. 2015	Benthic meiofauna from seagrass meadows	2mm sieve, 1mm sieve, 0.5mm sieve	Roche 454 GS FLX	COI 18S	450 710	USEARCH 6.1 (QIIIME 1.7)	UCLUST de novo (QIIME 1.7)	GenBank Silva 115
This study	Meiofauna from coarse shell sand and muddy benthic sediment	Siphoning 125µm, flotation (MgCl2) 125µm, flotation (H2O) 45µm/70µm	Paired-end Illumina Mi-Seq	COI 18S (V1-V2 regions)	313 364	UCHIME (part of USEARCH 6.1.) (QIIME 1.9.1)	CROP COI: 92-94% 18S: 95-97%	BOLD, SweBol and own databases for Nemertea, Acoela, Oligochaeta), Genbank Silva 111

In this study we used samples from muddy and sandy marine sediments to examine how results of metabarcoding based surveys of meiofaunal communities are impacted by three different meiofaunal extraction methods and three different primer pairs for COI and 18S. In order to validate the reliability of the metabarcoding approach, we compare the results obtained with traditional morphology-based taxonomic assignment for two test groups, Xenacoelomorpha and Nematoda, the latter previously shown to be the dominant taxon in meiofaunal communities in terms of number of OTUs (Fonseca et al. 2010).

Materials and Methods

Sampling

Samples were collected in two ecologically distinct locations along the west coast of Sweden in August 2014.

Hållö island samples: Coarse shell sand was sampled by dredging at 7-8m depth along the north-eastern side of Hållö island near Smögen, Sotenäs municipality, Västra Götalands county (N 58° 20.32-20.38', E 11° 12.73-12.68').

Gullmarn Fjord samples: Soft mud was collected using a Waren dredge at 53 m depth in the Gullmarn Fjord near Lysekil, Lysekil municipality, Västra Götalands county (N 58° 15.73', E 11°26.10').

Meiofaunal extraction

Hållö island. Hållö island samples were extracted in the lab using two different variations of the flotation (decanting and sieving) technique.

Flotation (freshwater): Freshwater was used to induce an osmotic shock in meiofaunal organisms and force them to detach from heavy sediment particles. 200 mL of sediment were placed in a large volume of fresh water and thoroughly mixed to suspend meiofauna

and lighter sediment particles. The supernatant was sieved through a 1000 µm sieve to separate the macrofaunal fraction, which was then discarded. The filtered sample was sieved again through a 45 µm sieve to collect meiofauna and discard fine organic particles. This procedure was repeated three times. Meiofauna was then rinsed with seawater from the sieve into large falcon tubes. Twelve sediment samples were processed, ten of them were fixed immediately in 96% ethanol for molecular analysis and stored at -20°C. The other two samples were first screened for live representatives of Xenacoelomorpha, and later preserved in 4% formaldehyde for morphology-based identification of nematodes.

Flotation (MgCl2 solution): A 7.2% solution of MgCl2 was used to anesthetize meiofauna. As above, twelve samples were processed in total, ten of them were decanted through 125 µm sieve and fixed immediately in 96% ethanol for molecular analysis and stored at -20°C, while two samples were decanted through a 125 µm sieve which was subsequently placed in a petri dish with seawater. After 30 minutes, the petri dish as well as the inside of the sieve were searched for Xenacoelomorpha using a stereo microscope. Afterwards they were preserved in 4% formaldehyde for morphology-based identification of nematodes.

Gullmarn Fjord. Meiofauna was extracted from the Gullmarn Fjord samples using two different methods: flotation and siphoning.

Flotation (freshwater): Freshwater was used to induce an osmotic shock in meiofaunal organisms. 2.4 L of sediment were placed in a large volume of freshwater, thoroughly mixed to suspend meiofauna and lighter sediment particles. The supernatant was sieved through a 1000 µm sieve in order to separate macrofauna, which was then discarded. The filtered sample was then sieved three times through a 70µm sieve to collect meiofauna and discard fine organic particles. Meiofauna was then rinsed with seawater from the sieve into a large container and equally divided between 12 falcon tubes. Six samples were fixed in 96% ethanol for molecular analysis and stored at -20°C. Six samples were screened for live representatives of Xenacoelomorpha, and preserved in 4% formaldehyde for morphology-based identification of nematodes.

Siphoning: A total volume of 12 L of sediment was processed as follows: an approximately 5 cm thick layer of mud was placed in a container and covered with 20 cm of seawater. The sediment was allowed to settle for 20 hours. Half of the sediment area was then siphoned through a 125 µm sieve, the residue in the sieve was immediately fixed in 96% ethanol, large macrofauna was manually removed, and the entire volume was split equally into six samples and placed at -20°C for subsequent molecular analysis. The remaining half of the area was similarly siphoned through a 125 µm sieve, the sieve contents were stored in sea water, large macrofauna manually removed, the entire volume split into six samples, which were screened for live representatives of Xenacoelomorpha, and preserved in 4% formaldehyde for morphology-based identification of nematodes.

Morphology-based identification

Xenacoelomorpha. Four samples from Hållö and 12 samples from Gullmarn Fjord were used for morphology-based assessment of the diversity of Xenacoelomorpha. All samples

were stored in seawater and searched for Xenacoelomorpha with a stereo microscope. All specimens found were immediately identified to the lowest taxonomic rank possible using a compound microscope equipped with DIC.

Nematoda. Two samples from each location/extraction method were used to assess nematode diversity using morphology-based identification. Samples from Hållö (flotation with fresh water and MgCl2) and Gullmarn Fjord (siphoning) were processed whole and samples from Gullmarn Fjord extracted using flotation with fresh water were subsampled by taking 1/10 of the entire sample. Formaldehyde–preserved samples were transferred to glycerin using Seinhorst's rapid method as modified by De Grisse (1969). Permanent nematode mounts on glass slides were prepared using the paraffin wax ring method. It is common practice to estimate the diversity of marine nematodes by counting a predetermined number (usually 100 or 200) of randomly picked nematodes per sample (Vincx 1996), which may not provide sufficiently detailed results for samples with high diversity. Therefore, all nematode specimens were counted and identified for each analyzed sample. All nematode specimens were identified to genus, and, when possible, to species level.

DNA extraction, library preparation and sequencing

DNA extraction. 30 samples were processed for total DNA extraction, twelve from the Gullmarn Fjord and eighteen from Hållö island, using 10g of sediment and the PowerMax® Soil DNA Isolation Kit (MO BIO Laboratories), according to manufacturer's instructions.

Primer design. Illumina MiSeq reagent v3. produces paired-end reads of 300bp in length, allowing a maximum marker length of 500bp when taking into account a 50 bp overlap. Universal COI primers available for the Metazoa amplify a 658bp region (Folmer et al. 1994), which is too long for most NGS applications.

Accordingly, primers amplifiying a 313 bp fragment of the mitochondrial cytochrome oxidase 1 (COI) gene were used, as described in Bourlat et al. 2016. The primers used for COI are modified from Leray et al.'s 'mini-barcode' COI primers (mlCOIintF-dgHCO2198; Leray et al. 2013) by adding the Illumina MiSeq overhang adapter sequences. The Leray et al. 'mini-barcode' primers have been shown to amplify up to 91% of metazoan diversity in a sample (Leray et al. 2013). In combination with Leray et al.'s mini barcode forward primer (mlCOIintF), we used Folmer et al.'s COI reverse primer (dgHCO2198; Folmer et al. 1994) as well as a reverse primer developed by Lobo et al., shown to enhance amplification of the COI region in a wide range of invertebrates (Lobo et al. 2013).

For the 18S region, Illumina overhang adapter sequences were appended to the primers from Fonseca et al. (SSU_FO4-SSU_R22; Fonseca et al. 2010), yielding a 364 bp fragment. These primers target a homologous region of the gene and flank a region that is highly divergent, corresponding to the V1-V2 region of the 18S gene (Lindeque et al. 2013, Fonseca et al. 2010).

Sequence overlap in the paired-end reads was calculated in Geneious Kearse et al. 2012. COI shows a sequence overlap of 230 bp and 18S shows an overlap of 190 bp.

All primer sequences used are shown in Table 2.

Table 2.

Primer sequences used in this study

Marker	Primer name	Illumina adapter overhang (regular font), with primer sequence (in bold)
COI Leray	mlCOIintF	5'-TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG**GGWACWGGWTGAACW GTWTAYCCYCC**-3'
	dgHCO2198	5'-GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG**TAAACTTCAGGGTGAC CAAARAAYCA**-3'
COI Lobo	mlCOIintF	5'-TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG**GGWACWGGWTGAACW GTWTAYCCYCC**-3'
	LoboR1	5'-GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG**TAAACYTCWGGRTGW CCRAARAAYCA**-3'
18S	SSU_FO4	5'-TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG**GCTTGTCTCAAAGATTA AGCC**-3'
	SSU_R22	5'-GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG**GCCTGCTGCCTTCCTT GGA**-3'

Illumina MiSeq library preparation using fusion primers. For Illumina MiSeq library preparation, we used a dual PCR amplification method as described in Bourlat et al. (2016). The first PCR, the amplicon PCR, uses amplicon specific primers including the Illumina adapter overhang, as described above. The second PCR, the index PCR, allows the incorporation of Illumina index adapters using a limited number of cycles (Bourlat et al. 2016).

Amplicon PCR. PCR amplifications of the COI and 18S regions were set up as follows. For a 50µl reaction volume, we used 5µl Pfu polymerase buffer (10x), 1µl dNTP mix (final concentration of each dNTP 200µM), 0.5 µl of each primer at 50 pm/µl, 2 µl DNA template (~10 ng), 0.5µl Pfu DNA polymerase (Promega) and 40.5µl of nuclease free water. Each DNA sample was amplified with the 3 primer pairs described above (COI Leray, COI Lobo and 18S). PCR cycling conditions were 2 min at 95°C (1 cycle); 1 min at 95°C, 45 s at 57° C, 2 min at 72°C (35 cycles); 10 min at 72°C (1 cycle). The PCR was checked on a 2% agarose gel. 20µl of each PCR reaction were then purified with Agencourt® AMPure® XP paramagnetic beads (Beckman Coulter), allowing size selection of PCR fragments by using different PCR product to bead ratios (Bourlat et al. 2016).

Index PCR. For dual indexing we used the Nextera XT index kit (96 indices, 384 samples, Illumina) according manufacturers' instructions. Dual indexing allows an increase in the multiplex level of sequencing per lane, so that more samples can be sequenced on the same flow cell (Fadrosh et al. 2014). It also eliminates cross-contamination between samples and the occurrence of mixed clusters on the flow cell (Kircher et al. 2012). The index PCR was set up as 50µl reactions using 5µl of cleaned up PCR amplicons, 5µl of

Nextera XT Index Primer i5, 5µl of Nextera XT Index Primer i7, 25µl of 2x KAPA HiFi HotStart ready mix (Kapa Biosystems) and 10µl of nuclease free water. PCR cycling conditions were: 3 min at 95°C (1 cycle); 30 s at 95°C, 30 s at 55°C, 30 s at 72°C (8 cycles); 5 min at 72°C (1 cycle). A bead purification was carried out after the index PCR with Agencourt® AMPure® XP magnetic beads (Beckman Coulter) using a ratio of 0.8, allowing the selection of fragments larger than 200 bp. DNA was quantified before sequencing using a Qubit Fluoremeter (Invitrogen) and average fragment size was verified using Tapestation (Agilent Technologies). Further library normalization and pooling steps are described in *Bourlat et al. (2016)*.

Sequencing. The pooled libraries were sequenced three times independently using Illumina MiSeq Reagent Kit v3, producing in total 24 132 875 paired-end reads of 300 bp in length, of which 15 883 274 COI reads and 8 249 601 18S reads (Table 3).

Table 3.

Number of reads per marker and per sequencing run

Marker / Sequencing run	1	2	3	Total
COI	5 859 454	5 075 735	4 948 085	15 883 274
18S	2 803 391	3 135 331	2 310 879	8 249 601
Total	8 662 845	8 211 066	7 258 964	24 132 875

Bioinformatic data processing and analysis

Most analytical steps were performed using Qiime (Quantitative Insight Into Microbial Ecology) version 1.9.1 (Caporaso et al. 2010) and custom python scripts (Fig. 1).

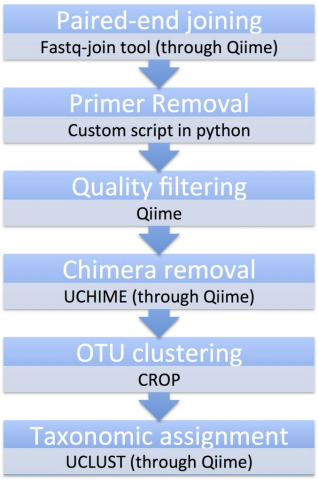


Figure 1. doi

Schematic workflow of bioinformatic analytical steps

Paired-end joining

Demultiplexed MiSeq paired-end reads were joined using the Qiime script *multiple_join_paired_ends.py* using the fastq-join tool (https://code.google.com/p/ea-utils/wiki/FastqJoin). Data from three sequencing runs were merged producing a total of 24 132 875 raw paired-end reads, 15 883 274 reads for the COI dataset and 8 249 601 reads for the 18S dataset (Table 3). The number of reads remaining after various bioinformatic data processing steps is presented in Table 4. After paired-end joining, 48% of sequences were lost leading to a total of 12 543 198 reads, due to an observed decrease in sequence quality at the end of the reads, resulting in a bad overlap between the paired-ends. This loss is much more important for the longer 18S region (2 131 102 reads after joining, corresponding to a 74% loss) than for the COI region (10 412 096 reads after joining, corresponding to a 34,5% loss).

Table 4.

Number of reads remaining after each bioinformatic step

Marker / Step	Raw data	Paired-end joining	Primer trimming	Quality filtering	Chimera removal
COI	15 883 274	10 412 096	8 099 507	7 976 649	7 954 017
18S	8 249 601	2 131 102	1 071 871	1 015 874	890 370
Total	24 132 875	12 543 198	9 171 378	8 992 523	8 844 387

Primer trimming and quality filtering

Dual indexes and Illumina overhangs were removed by the sequencing platform. COI and 18s primer sequences were removed using a custom python script designed for this study (https://github.com/Quiterie90/Primer_Removal). The script retains and trims reads that have the exact sequence of the forward and reverse primers at the beginning and at the end of the reads respectively, while other reads not meeting these criteria are discarded. The script takes into account the presence of ambiguous bases in the primer sequence (such as W, R, S, Y, M, K, H, D, B and V). In the case that an unassigned base (N) is found in the primer sequence, the read is also discarded. The primer-trimming step resulted in 9 171 378 reads remaining corresponding to a 27% loss. As the script is quite stringent, it quality filters reads by removing incomplete reads or chimeras. At this step 1 071 871 reads remained after trimming for the 18S dataset corresponding to a 50% loss and 8 099 507 reads remained after trimming for the COI dataset corresponding to a 22% loss. A quality filtering step was then carried out using the Qiime script *multiple_split_libraries_fastq.py* to remove reads with a Q Score inferior to 30 (corresponding to a base call accuracy of at least 99,9%). A total of 2% of sequences were lost after the quality-filtering step leading to 8 992 523 reads remaining. 5% of the reads were lost in the 18S dataset corresponding to a final 1 015 874 reads and 1,5% of the reads were lost in the COI dataset corresponding to a final 7 976 649 reads.

Chimera removal and OTU clustering

Chimeric reads were removed with UCHIME (Edgar et al. 2011) using the Qiime scripts *identify_chimeric_seqs.py* followed by *filter_fasta.py* based on the Usearch61 software. After chimera removal, 7 954 017 sequences remained in the COI dataset (0,3 % loss) and 890 370 sequences remained in the 18S dataset (12% loss).

For clustering sequences into Operational Taxonomic Units (OTUs) we used CROP, a Bayesian clustering algorithm that delineates OTUs based on the natural distribution of the data, using a Gaussian mixture model (Hao et al. 2011). The program allows the user to define a lower and upper bound variance to cluster the sequences, instead of a fixed sequence similarity value. According to a benchmarking study by Leray *et al.* based on the Moorea Biocode barcode library (http://mooreabiocode.org/; Leray et al. 2013), the best lower and upper bound values to cluster metazoan COI sequences are 3 and 4, corresponding to sequence dissimilarities between 6% and 8%. According to an 18S benchmarking experiment with a set of 41 known nematode species carried out by Porazinska *et al.*, a 96% threshold most accurately reflects taxonomic richness, yielding 37 OCTUs, whereas a 97% threshold yielded 51 OCTUs (Porazinska et al. 2009). According to this benchmark, a range of sequence dissimilarities between 3% and 5% were used in CROP (1.5 and 2.5 respectively for the lower and upper values, corresponding to 95-97% similarity).

Parameters used in CROP for the analysis were as follows:

CROP -i -b 160 000 -z 470 -l 3 -u 4 -o

CROP -i -b 18 000 -z 470 -l 1.5-u 2.5 -o

The 7 954 017 COI sequences and the 890 370 18S sequences were clustered into 2805 and 1472 representative OTUs respectively, 213 of which were identified to species for COI and 243 of which were identified to species for 18S, using a 97% sequence similarity threshold (Table 5 Fig. 2).

Table 5.

Number of OTUs and percentage per phylum for COI and 18S for the metazoan fraction. Based on a 97% similarity threshold.

Phylum	COI		18S	
	OTUs	Percentage	OTUs	Percentage
Annelida	57	30.00	29	23.97
Arthropoda	52	27.37	14	11.57
Bryozoa	5	2.63	3	2.48
Cephalorhyncha	0	0.00	1	0.83
Chaetognatha	1	0.53	0	0.00
Chordata	12	6.32	7	5.79

Cnidaria	8	4.21	4	3.31
Echinodermata	13	6.84	5	4.13
Gastrotricha	1	0.53	9	7.44
Gnathostomulida	1	0.53	0	0.00
Mollusca	26	13.68	6	4.96
Nematoda	0	0.00	10	8.26
Nemertea	3	1.58	6	4.96
Platyhelminthes	0	0.00	13	10.74
Phoronida	1	0.53	0	0.00
Porifera	2	1.05	3	2.48
Priapulida	1	0.53	0	0.00
Rotifera	2	1.05	0	0.00
Sipuncula	1	0.53	1	0.83
Tardigrada	0	0.00	1	0.83
Xenacoelomorpha	4	2.11	9	7.44
Total OTUs Metazoa	190	100	121	100

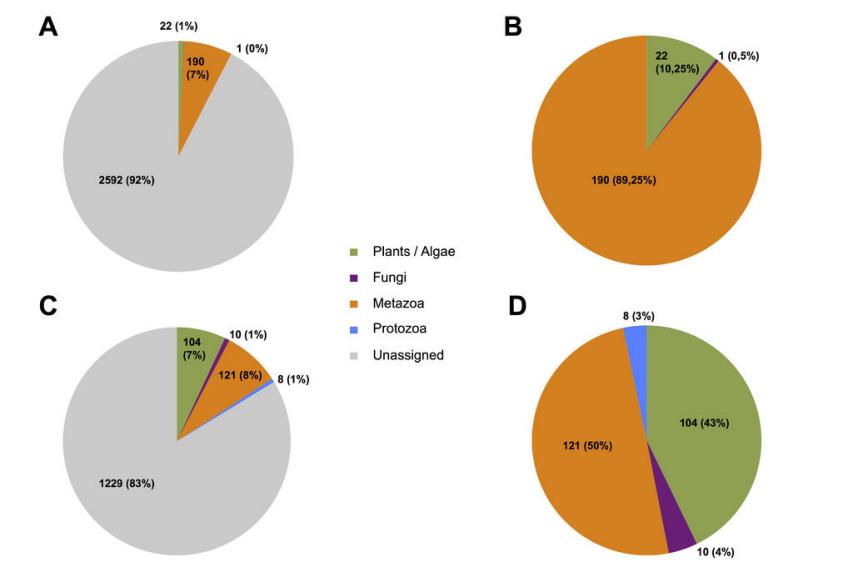


Figure 2. doi

Taxonomic composition overview at species level based on a 97% sequence similarity threshold. A) Percentages and counts of OTUs for the COI gene with unassigned OTUs. B) Percentages and counts of OTUs for the COI gene without unassigned OTUs. C) Percentages and counts of OTUs for the 18S gene with unassigned OTUs. D) Percentages and counts of OTUs for the 18S gene without unassigned OTUs.

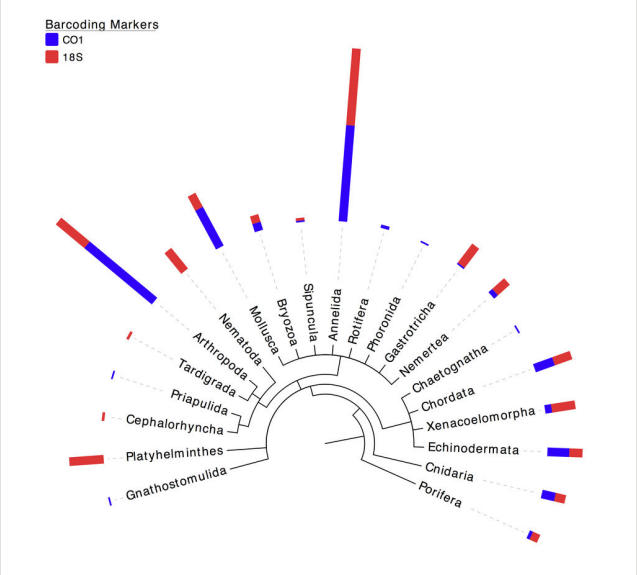


Figure 3. doi

Percentages of metazoan phyla uncovered in the samples using COI and 18S molecular surveys. Blue bars correspond to the cumulated frequencies of OTUs assigned to a specific phylum using the COI gene and red bars correspond to the cumulated frequencies of OTUs assigned to a specific phylum using the 18S gene. Taxonomic assignment is based on a 97% sequence similarity threshold.

Taxonomic assignment

As Qiime is normally used for metagenomic analyses of prokaryotes, default databases are not suited for taxonomic assignment of Metazoa. Custom databases consisting in a taxonomy file associated with a reference sequence file can be created, or alternatively, a preformatted database such as the Silva database (http://www.arb-silva.de/no_cache/download/archive/qiime/) can be used. For the COI region, a custom database of 1 947 954 sequences was created consisting of the BOLD database (http://www.boldsystems.org/ downloaded on October 8 2015), combined with own reference databases of Nemertea, Xenacoelomorpha and Oligochaeta and barcodes of Swedish Echinodermata, Mollusca, Cnidaria and Arthropoda from the Swedish Barcode of Life database (SweBol). For the 18S rRNA region, a custom database of 732 419 reference sequences was created using the Silva database release 111 (http://www.arb-silva.de/no_cache/download/archive/qiime/) and own barcodes for Acoela and Oligochaeta. Corresponding tab-delimited taxonomy files were created including a sequence ID and taxonomic lineage information (Phylum, Class, Order, Family, Genus and Species) derived from BOLD, Swebol, Silva and WoRMS (http://www.marinespecies.org/).

Taxonomic assignments were carried out using both 80% and 97% sequence similarity thresholds, to obtain identifications at phylum and species levels respectively (Giongo et al.

2010, Lanzén et al. 2012), yielding 690 metazoan OTUs for COI and 793 metazoan OTUs for 18S at 80% threshold and 190 metazoan OTUs for COI and 121 metazoan OTUs for 18S at 97% threshold. For COI, taxonomic assignment was done with the Qiime script *assign_taxonomy.py* using the Uclust software (Edgar 2010). With Uclust, a query sequence matches a database sequence if the identity is high enough. The identity is calculated from a global alignment, which differs from BLAST and most other database search programs, which search for local matches. By default, Uclust stops searching when it finds a match, but also stops searching if it fails to find a match after eight failed attempts. Within Qiime, Uclust is the default algorithm for the *assign_taxonomy.py* script and two parameters are associated to the algorithm. The minimum fraction of database hits that must have a specific taxonomic assignment to assign that taxonomy to a query that was fixed at 0.51 and the number of database hits to consider when making an assignment that was fixed at 3, corresponding to the default values. To obtain matches for non-Metazoan taxa, a Megablast search with 70% minimum coverage was done against the Genbank nt (nucleotide) database (ftp://ftp.ncbi.nlm.nih.gov/blast/db/ downloaded on June 27 2015) using Geneious (Kearse et al. 2012). For taxonomic assignment of the 18S dataset, the Qiime script *assign_taxonomy.py* was used with Uclust (Edgar 2010) default settings against the Silva database. Some taxonomic errors were detected for Nematodes in the Silva database.

Note on the taxonomic assignement of Nematodes: The output from the Qiime analysis included 145 18S OTUs assigned to the phylum Nematoda. Three of them (HE1.SSU866120, HE6.SSU382930 and HF6.SSU331569) Suppl. material 1 were incorrectly placed among the nematodes due to errors in the reference database they derived from – they group among Arthropod taxa by the Megablast search and were excluded for that reason. Another OTU (TS6.SSU559982) is placed among Phoronida by the Megablast search and was also excluded. Two more sequences that were assigned to Nematoda appear to have long insertions within conserved regions (HE6.SSU358113 and TF5.SSU411806). Both of them were found only in one sample each, further supporting the idea that they are derived from erroneous amplification product, and were removed from any further analysis.

Invasive Alien Species (IAS) were detected in our samples by comparing our species list (Suppl. material 1) to the Helcom-Ospar list (http://www.helcom.fi/about-us/partners/ospar) and the Swedish Främmande Arter invasive species lists (http://www.frammandearter.se/).

Taxonomic composition bar plots (Fig. 4) were created using OTU tables (Suppl. materials 2, 3) and the Qiime scripts *make_otu_table.py*, *split_otu_table_by_taxonomy*, *merge_otu_table.py* and *summarize_taxa_through_plots.py*. The bar plots created for Fig. 4 take into account the relative abundance or number of reads for each OTU, whereas Table 5 and Fig. 3 do not take relative abundances of each OTU into account. Fig. 3 showing community composition per phylum and marker was created using PhyloT (http://phylot.biobyte.de/) and Evolview tools (http://www.evolgenius.info/evolview.html; (Zhang et al. 2012).

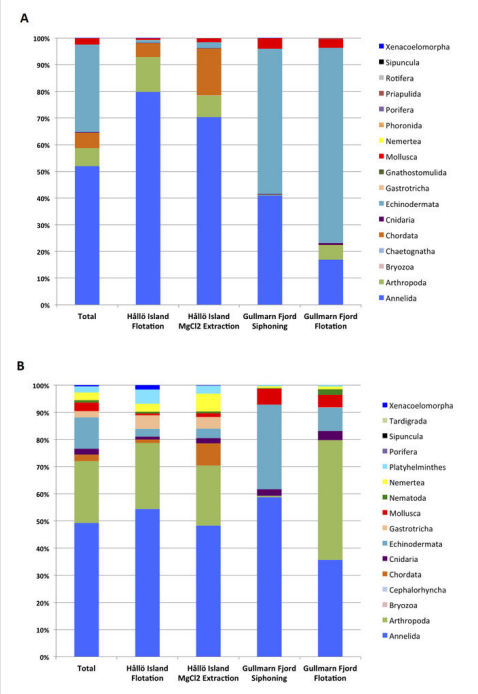


Figure 4. doi

Community composition per phylum in Hållö island and Gullmarn fjord samples, according to extraction method (MgCl2, H2O, Siphoning). A) For the COI gene. B) For the 18S gene. The vertical axis corresponds to percentage of OTUs. Taxonomic assignment is based on a 97% similarity threshold. The bar plots take into account number of reads for each OTU.

Diversity analyses

Alpha and beta diversity analyses were carried out with and without unassigned OTUs for both COI and 18S datasets. Unassigned OTUs were removed using the Qiime script *filter_otus_from_otu_table.py*. Alpha diversity (species richness) was calculated using the nonparametric Chao1 index using rarefied datasets to correct bias in species number due to unequal sample size. One of the samples in the COI dataset was removed prior to rarefaction analysis due to low sequence number (1122 sequences including unassigned OTUs and 280 sequences excluding unassigned OTUs at 97% sequence similarity) using the Qiime script *filter_sample_from_otu_table.py*. Rarefaction, alpha diversity calculation and generation of plots were performed using the Qiime scripts i) *multiple_rarefactions.py,* ii) *alpha_diversity.py*, iii) *collate_alpha.py* and iv) *make_rarefaction_plots.py*. Rarefaction was done to a depth corresponding to the total number of sequences in the smallest dataset (20405 sequences including unassigned OTUs and 5442 sequences excluding unassigned OTUs at 97% sequence similarity for COI, and 7561 sequences including unassigned OTUs and 5399 sequences excluding unassigned OTUs at 97% sequence similarity for 18S). Alpha diversities were compared between locations and extraction

methods for both datasets and COI primer sets using the Qiime script *compare_alpha_diversity.py*. The script performs Monte-Carlo permutations to determine p-values.

Beta diversity was calculated using the abundance-based Bray-Curtis index for both COI and 18S datasets. The Qiime script *beta_diversity_through_plots.py* was used to compute beta diversity distance matrices from the rarefied samples and generate Principal Coordinate Analysis (PCoA) plots. Beta diversity was compared according to location, extraction method and primer pair both with and without the unassigned OTUs using the Qiime script *compare_categories.py*. The script uses R and the vegan and ape libraries to compute statistical tests. We performed ANOSIM (ANalysis Of SIMilarity) tests, which are nonparametric, through 999 permutations. This method tests whether two or more groups of samples are significantly different by taking as null hypothesis that there is no difference between the two or more groups studied.

Alpha and beta diversities were calculated including and excluding the unassigned OTUs and results obtained were similar. Here we present plots including the unassigned OTUs (Figs 5, 6).

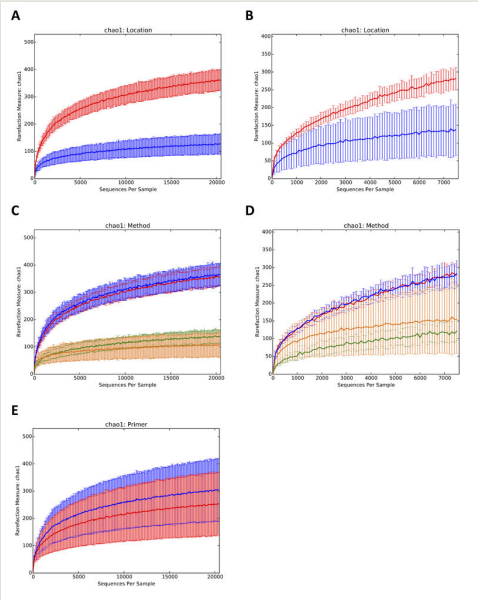


Figure 5. doi

Alpha diversity rarefaction plots for COI and 18S datasets including unassigned OTUs. According to location for COI (A) 18S (B). Hållö Island (HI) in red, Gullmarn Fjord (GF) in blue. According to extraction method for COI (C) 18S (D). HI flotation in red, HI MgCl2 in blue, GF flotation in yellow, GF siphoning in green. According to primer pair for COI (E). CO1 Leray primer in red, COI Lobo primer in blue.

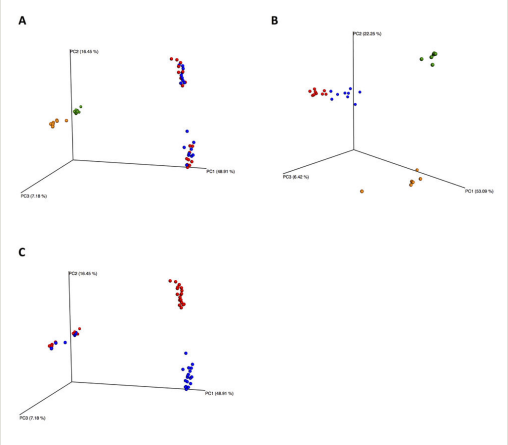


Figure 6. doi

Beta diversity PCoA plots for COI and 18S datasets including unassigned OTUs. According to extraction method for COI (A) 18S (B) HI flotation in red, HI MgCl2 in blue, GF flotation in yellow and GF siphoning in green. According to primer for COI (C) COI Leray primer in red, COI Lobo primer in blue

Data resources

The data underpinning the analysis reported in this paper are deposited at the GenBank SRA under project number PRJNA388326 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA388326).

Results and discussion

Phylum-level community composition of meiofaunal samples from the Swedish west coast

Illumina MiSeq produced at total of 24 132 875 raw reads, of which 15 883 274 COI reads and 8 249 601 18S reads. These were quality filtered (see methods section for details) resulting in 7 954 017 COI sequences and 890 370 18S sequences. These were clustered into 2805 and 1472 representative OTUs respectively, yielding 190 metazoan OTUs for COI and 121 metazoan OTUs for 18S at 97% sequence similarity (see methods, Table 5 & Fig. 2).

Taxonomic assignment of OTUs at a 97% similarity threshold shows community composition of the samples at the phylum level (Fig. 2). Of 2805 COI OTUs, 190 (7%) were assigned to the Metazoa, 22 (1%) to plants and algae, 1 (0%) to Fungi. 2592 OTUs remained unassigned, corresponding to 92% of COI OTUs.

For the 18S dataset, 121 of 1472 OTUs (8%) were assigned to Metazoa, 104 (7%) to plants and algae, 10 (1%) to Fungi, and 8 (1%) to Protozoa. 1229 OTUs remained unassigned, corresponding to 83% of all 18S OTUs.

The large numbers of unassigned OTUs reflect the incompleteness of the databases used for COI and 18S. When unassigned OTUs are disregarded, differences between the taxonomic ocverage of the markers can be observed (Fig. 2, B and D). COI is the 'standard' animal barcode and is thus mostly useful for diversity surveys within the Metazoa (Hebert et al. 2003). 18S has on the other hand much larger taxonomic coverage and can be used for biodiversity profiles of whole eukaryotic communities, at higher taxonomic scales.

Of all OTUs classified as Metazoa, a detailed breakdown per phylum is presented in Table 5 and Fig. 3. Annelida (30% of CO1 metazoan OTUs and 23.97% of 18S metazoan OTUs) and Arthropoda (27.37% of CO1 metazoan OTUs and 11.57% of 18S metazoan OTUs), were the most OTU rich phyla identified in all samples combined, a similar pattern as observed in a recent study on coastal seagrass meadows in Brittany, France (Cowart et al. 2015).

As well as Annelida and Arthropoda, other phyla represented by a high number of OTUs in our samples include Mollusca (13.68% of COI metazoan OTUs and 4.96% of 18S metazoan OTUs), Platyhelminthes (10,74% of 18S metazoan OTUs and 0% of CO1 metazoan OTUs) and Nematoda (8.26% of 18S metazoan OTUs and 0% of CO1 metazoan OTUs) (Table 5 & Fig. 3). Other benthic metabarcoding studies based on the 18S V1-V2 region, found Nematoda and Platyhelminthes as the most OTU rich phyla represented (Fonseca et al. 2014, Fonseca et al. 2010), or Nematoda and Annelida (Bik et al. 2012b), alternatively Nematoda and Arthropoda (Bik et al. 2012a, Lallias et al. 2015).

Meiofaunal community composition differs according to location

Taxonomic community composition at both locations surveyed is illustrated in Fig. 4. The bar plots in Fig. 4 take into account the read counts for each OTU, whereas Table 5 and Fig. 3 do not take these into account.

In Fig. 4, clear differentiation in biodiversity between the two habitat types (soft mud versus coarse shell sand) can be observed, as expected. Echinodermata (such as Ophiurida, Echinoidea and Asteroidea), Mollusca (Bivalvia, Gastropoda), Annelida and Arthropoda are represented by higher numbers of reads in samples from the muddy sediments in the Gullmarn fjord samples (grain size 100 µm approx.).

In coarse shell sand in shallow areas, such as in the Hållö island samples, Annelida and Arthropoda are represented by higher numbers of reads, followed by Chordata (cephalohordata such as *Branchiostoma* sp., ascidians and various fish species such as *Gobius* sp., *Ctenolabrus rupestris, Solea solea*) with in addition a larger diversity of small taxa such as Bryozoa, Gnathosthomulida, Gastrotricha, Tardigrada, Rotifera, Sipuncula and Phoronida, reflecting the high diversity of insterstitial taxa found in sandy sediments.

Sample diversity and composition analyses

A greater number of phyla were uncovered in the Hållö Island samples than in the Gullmarn Fjord samples (Fig. 4A and 4B) and this observation was corroborated by the alpha diversity rarefaction plots showing that Hållö Island samples (in red) present a higher diversity than the Gullmarn Fjord samples (in blue) (p-value = 0.001) regardless of the marker used (Fig. 5A and 5B). Within the same location, choice of extraction method does not have a significant impact on sample diversity (p-value ~ 1) (Fig. 5C and 5D, Table 6). However, for the 18S dataset, the flotation method seems to be more effective for extraction of nematodes than the siphoning method in the Gullmarn Fjord samples (Fig. 4A and 4B). Moreover, the beta diversity PCoA results highlight the fact that sample composition is influenced by the choice of extraction method for both COI and 18S datasets (p-value = 0.001) leading to four different clusters (Fig. 6and 6B, Table 6). For the COI dataset, in addition to extraction method as a factor of divergence, choice of primer (COI Leray or COI Lobo) also influences the grouping of the samples (p-value = 0.003 excluding unassigned OTUs and 0.001 including unassigned OTUs), in particular for the Hållö Island samples (Fig. 6C). Moreover, the COI Lobo primer seems to uncover a higher diversity of taxa than the COI Leray primer Fig. 5E) even if the results are considered to be non significant (p-value = 0.585 excluding unassigned OTUs and 0.111 including unassigned OTUs) (Table 6Table 7).

Table 6.

Nonparametric t-test results with 999 Monte-Carlo permutations for both datasets with and without unassigned OTUs (97% taxonomic assignment)

	COI dataset				**18S dataset**			
	Excluding unassigned OTUs		Including Unassigned OTUs		Excluding unassigned OTUs		Including Unassigned OTUs	
	Test value	P-value	Test value	P-value	Test value	P-value	Test value	P-value
Location								
HI vs. GF	-14.453	0.001	-21.455	0.001	-6.929	0.001	-7.170	0.001
Method								
HI H2O vs. HI MgCl2	-0.437	1.0	-0.691	1.0	-0.906	1.0	-0.174	1.0
GF flotation vs. GF siphoning	1.567	0.792	1.546	0.99	-1.427	1.0	-0.744	1.0
Primer								
COI Leray vs. COI Lobo	-0.508	0.596	-1.614	0.111	-	-	-	-

Table 7.

ANOSIM test results (999 permutations) for both COI and 18S datasets with and without unassigned OTUs (97% taxonomic assignment)

	COI dataset				**18S dataset**			
Ho: Sample composition differs according to	Excluding unassigned OTUs		Including unassigned OTUs		Excluding unassigned OTUs		Including unassigned OTUs	
	R-value	P-value	R-value	P-value	R-value	P-value	R-value	P-value
Location	0.976	0.001	1.0	0.001	0.935	0.001	0.929	0.001
Method	0.660	0.001	0.738	0.001	0.889	0.001	0.895	0.001
Primer	0.200	0.003	0.218	0.001	-	-	-	-

Molecular identifications to species level

Using a sequence similarity search at 97% similarity allowed us to identify 213 COI OTUs and 243 18S OTUs to species level (Table 8 and Suppl. material 1). For the COI dataset, 81 species (of which 70 metazoans) were found in both locations, 36 (of which 35 metazoans) were found in the Gullmarn fjord only and 96 (of which 85 metazoans) were found in Hållö island only. For the 18S dataset, 108 species (of which 48 metazoans) were found in both locations, 44 (of which 21 metazoans) were found in the Gullmarn fjord only and 91 (of which 52 metazoans) were found in Hållö Island only (Suppl. material 1). These species observations from metabarcoding represent 'molecular occurrence records' that could be used in monitoring and other types of biodiversity surveys, in the same way as physical observations, such as for mapping species distributions (Bohmann et al. 2014, Lawson Handley 2015).

Table 8.

Metazoa identified to species level using 97% sequence similarity (HI: Hållö island, GF: Gullmarn Fjord)

COI							
OTU ID	Nb of reads	Phylum	Class	Order	Species	HI	GF
HE6.Lobo_7972794	3	Annelida	Clitellata	Haplotaxida	*Adelodrilus pusillus*	+	-
HE1.Lobo_933012	14954	Annelida	Clitellata	Haplotaxida	*Grania postclitellochaeta*	+	+
HF8.Lobo_5239705	241	Annelida	Clitellata	Haplotaxida	*Grania variochaeta*	+	+
HF4.Lobo_97092	29391	Annelida	Clitellata	Haplotaxida	*Tubificoides benedii*	+	+

HF5.Lobo_3297996	1	Annelida	Clitellata	Haplotaxida	*Tubificoides kozloffi*	+	-
TS1.Leray_545620	7370	Annelida	Polychaeta	Amphinomida	*Paramphinome jeffreysii*	-	+
HF1.Lobo_4996219	4596	Annelida	Polychaeta	Canalipalpata	*Polygordius appendiculatus*	+	+
TF6.Lobo_5247622	9030	Annelida	Polychaeta	Capitellida		-	+
TS1.Lobo_4669404	5	Annelida	Polychaeta	Capitellida		-	+
TF5.Lobo_6394093	2	Annelida	Polychaeta	Capitellida		-	+
TS3.Leray_6813257	1852	Annelida	Polychaeta	Eunicida		-	+
HF5.Leray_4035802	1	Annelida	Polychaeta	Eunicida	*Ophryotrocha maculata*	+	-
TS2.Leray_4445240	8815	Annelida	Polychaeta	Eunicida	*Parougia eliasoni*	+	+
TF3.Leray_6645504	5196	Annelida	Polychaeta	Opheliida		+	+
TS5.Lobo_6031643	5089	Annelida	Polychaeta	Opheliida		+	+
HF9.Lobo_7587930	1	Annelida	Polychaeta	Opheliida		+	-
HE8.Leray_7284535	2	Annelida	Polychaeta	Phyllodocida		+	-
TS5.Leray_1557252	88	Annelida	Polychaeta	Phyllodocida		-	+
TS3.Leray_6744085	1	Annelida	Polychaeta	Phyllodocida		-	+
TS3.Leray_6805306	2	Annelida	Polychaeta	Phyllodocida	*Aphrodita aculeata*	-	+
TS3.Lobo_1308935	4213	Annelida	Polychaeta	Phyllodocida	*Eumida ockelmanni*	+	+
HE6.Leray_2958692	69642	Annelida	Polychaeta	Phyllodocida	*Glycera alba*	+	+
HF7.Leray_1672792	69	Annelida	Polychaeta	Phyllodocida	*Glycinde nordmanni*	+	+
TF5.Leray_2872180	7754	Annelida	Polychaeta	Phyllodocida	*Gyptis mackiei*	-	+
HF1.Lobo_5059232	13	Annelida	Polychaeta	Phyllodocida	*Gyptis propinqua*	+	-
HF9.Lobo_7695035	1	Annelida	Polychaeta	Phyllodocida	*Lepidonotus squamatus*	+	-
HE6.Lobo_7972042	2	Annelida	Polychaeta	Phyllodocida	*Myrianida edwarsi*	+	-
HF9.Lobo_7688887	3	Annelida	Polychaeta	Phyllodocida	*Nereimyra punctata*	+	-
HF2.Lobo_2136301	178929	Annelida	Polychaeta	Phyllodocida	*Pisione remota*	+	+

HE3.Leray_364663	59407	Annelida	Polychaeta	Phyllodocida	*Platynereis dumerilli*	+	+
TS4.Leray_7471107	1	Annelida	Polychaeta	Phyllodocida	*Sige fusigera*	-	+
HE5.Lobo_493462	571790	Annelida	Polychaeta			+	+
TS2.Lobo_6962270	4595	Annelida	Polychaeta	Sabellida	*Galathowenia oculata*	+	+
TS2.Leray_4491798	316559	Annelida	Polychaeta	Spionida		+	+
TS4.Lobo_1502925	195999	Annelida	Polychaeta	Spionida		+	+
HF9.Lobo_7588557	891	Annelida	Polychaeta	Spionida		+	-
TS6.Leray_5665274	936	Annelida	Polychaeta	Spionida		-	+
TF1.Lobo_2668551	874	Annelida	Polychaeta	Spionida		-	+
HE4.Leray_3067470	3	Annelida	Polychaeta	Spionida	*Chaetopterus sarsi*	+	-
HF1.Lobo_4965916	1	Annelida	Polychaeta	Spionida	*Malacoceros fuliginosus*	+	-
HF9.Leray_4404528	1	Annelida	Polychaeta	Spionida	*Polydora cornuta*	+	-
HF5.Lobo_3178682	2894	Annelida	Polychaeta	Spionida	*Spiophanes bombyx*	+	+
TF1.Leray_2314881	29235	Annelida	Polychaeta	Terebellida		+	+
TF1.Lobo_2832834	9348	Annelida	Polychaeta	Terebellida		+	+
TS1.Leray_614419	788	Annelida	Polychaeta	Terebellida		+	+
HE8.Lobo_858951	1	Annelida	Polychaeta	Terebellida		+	-
TS2.Lobo_6889557	184	Annelida	Polychaeta	Terebellida		-	+
TS6.Lobo_255019	3	Annelida	Polychaeta	Terebellida		-	+
TS2.Lobo_6860909	1	Annelida	Polychaeta	Terebellida		-	+
TS5.Leray_1638640	1	Annelida	Polychaeta	Terebellida		-	+
TF1.Lobo_2848745	1305	Annelida	Polychaeta	Terebellida	*Amphictene auricoma*	+	+
TS3.Leray_6729893	1	Annelida	Polychaeta	Terebellida	*Brada villosa*	-	+
HF4.Lobo_96799	102	Annelida	Polychaeta	Terebellida	*Cirratulus cirratus*	+	-
HF2.Lobo_2052205	285	Annelida	Polychaeta	Terebellida	*Dodecaceria concharum*	+	-
TS5.Leray_1638834	102	Annelida	Polychaeta	Terebellida	*Lagis koreni*	+	+

HE9.Lobo_2191024	8	Annelida	Polychaeta	Terebellida	*Macrochaeta clavicornis*	+	+
TF1.Leray_2475372	6353	Annelida	Polychaeta	Terebellida	*Sosane wahrbergi*	+	+
HE1.Lobo_982378	38	Arthropoda	Branchiopoda	Diplostraca	*Evadne nordmanni*	+	-
TF5.Lobo_6391642	10097	Arthropoda	Branchiopoda	Diplostraca	*Penilia avirostris*	+	+
HF9.Lobo_7623741	1	Arthropoda	Branchiopoda	Diplostraca	*Pleopis polyphemoides*	+	-
TS4.Leray_7402581	10	Arthropoda	Insecta	Diptera		+	+
TS3.Lobo_1162454	2	Arthropoda	Insecta	Diptera	*Chironomus aprilinus*	+	+
HF4.Lobo_5006	1	Arthropoda	Insecta	Diptera	*Cryptochironomus supplicans*	+	-
TF5.Leray_2910679	6	Arthropoda	Insecta	Diptera	*Procladius* sp.	+	+
HF9.Lobo_7599310	3	Arthropoda	Insecta	Diptera	*Psectrocladius yunoquartus*	+	+
HE5.Lobo_479906	152	Arthropoda	Insecta	Diptera	*Tanytarsus usmaensis*	+	+
HE2.Lobo_2023271	21589	Arthropoda	Malacostraca	Amphipoda		+	+
HF1.Leray_2493444	3911	Arthropoda	Malacostraca	Amphipoda		+	-
HE8.Lobo_860608	1	Arthropoda	Malacostraca	Amphipoda		+	-
HE3.Lobo_4900763	1	Arthropoda	Malacostraca	Amphipoda	*Ampelisca brevicornis*	+	-
HF4.Leray_6193380	66039	Arthropoda	Malacostraca	Amphipoda	*Atylus vedlomensis*	+	+
HE8.Leray_7216397	1	Arthropoda	Malacostraca	Amphipoda	*Corophium volutator*	+	-
HE6.Lobo_7849183	1	Arthropoda	Malacostraca	Amphipoda	*Leptocheirus hirsutimanus*	+	-
HE1.Lobo_914374	14588	Arthropoda	Malacostraca	Amphipoda	*Monocorophium insidiosum*	+	+
TF1.Leray_2445583	56	Arthropoda	Malacostraca	Amphipoda	*Monoculodes packardi*	-	+
TF6.Leray_5321299	11588	Arthropoda	Malacostraca	Cumacea		+	+
HF9.Leray_4291607	1372	Arthropoda	Malacostraca	Decapoda	*Athanas nitescens*	+	-

HF8.Leray_5586003	2864	Arthropoda	Malacostraca	Decapoda	*Eualus cranchii*	+	+
HF8.Leray_5612792	37	Arthropoda	Malacostraca	Decapoda	*Eualus cranchii*	+	-
HE1.Lobo_952576	3739	Arthropoda	Malacostraca	Decapoda	*Liocarcinus navigator*	+	-
TF5.Lobo_6459477	1279	Arthropoda	Malacostraca	Decapoda	*Philocheras bispinosus bispinosus*	+	+
HE4.Lobo_4138563	42	Arthropoda	Malacostraca	Decapoda	*Pisidia longicornis*	+	+
HE8.Leray_7306131	2	Arthropoda	Malacostraca	Decapoda	*Processa modica*	+	-
TS3.Lobo_1213146	17	Arthropoda	Malacostraca	Isopoda	*Asellus aquaticus*	+	+
TF5.Leray_2897128	3	Arthropoda	Maxillopoda	Calanoida	*Acartia bifilosa*	-	+
HF3.Leray_7129076	22	Arthropoda	Maxillopoda	Calanoida	*Acartia clausi*	+	+
TF6.Leray_5332240	7399	Arthropoda	Maxillopoda	Calanoida	*Acartia tonsa*	+	+
HF7.Leray_1683272	927	Arthropoda	Maxillopoda	Calanoida	*Acartia tonsa*	+	+
HE2.Lobo_2010882	1	Arthropoda	Maxillopoda	Calanoida	*Anomalocera patersoni*	+	-
TS2.Leray_4478240	2	Arthropoda	Maxillopoda	Calanoida	*Calanus euxinus*	-	+
HF7.Lobo_5810493	41	Arthropoda	Maxillopoda	Calanoida	*Centropages hamatus*	+	+
HF8.Lobo_5106754	82	Arthropoda	Maxillopoda	Calanoida	*Centropages typicus*	+	+
HE8.Leray_7251655	1	Arthropoda	Maxillopoda	Calanoida	*Eurytemora affinis*	+	-
HE7.Leray_3803390	5325	Arthropoda	Maxillopoda	Calanoida	*Paracalanus parvus*	+	+
HF9.Leray_4411242	1	Arthropoda	Maxillopoda	Calanoida	*Pseudocalanus elongatus*	+	-
TS4.Leray_7515925	2	Arthropoda	Maxillopoda	Calanoida	*Pseudocalanus elongatus*	-	+
TS3.Lobo_1208165	1	Arthropoda	Maxillopoda	Calanoida	*Scolecithricella minor*	-	+
TF5.Lobo_6373065	809	Arthropoda	Maxillopoda	Calanoida	*Temora longicornis*	+	+
TF1.Leray_2453024	1	Arthropoda	Maxillopoda	Calanoida	*Temora longicornis*	-	+
HF4.Leray_6242499	45	Arthropoda	Maxillopoda	Cyclopoida		+	-

HF4.Leray_6206299	2	Arthropoda	Maxillopoda	*Harpacticoida*		+	-
HE8.Lobo_823478	108	Arthropoda	Maxillopoda	*Harpacticoida*	*Harpacticoida* sp.	+	-
TS3.Lobo_1208905	116	Arthropoda	Maxillopoda	*Harpacticoida*	*Harpacticus flexus*	+	+
HE1.Lobo_995710	1	Arthropoda	Maxillopoda	*Harpacticoida*	*Tachidius discipes*	+	-
HF4.Leray_6092514	1	Arthropoda	Maxillopoda	Poecilostomatoida		+	-
HF9.Leray_4391714	11307	Arthropoda	Maxillopoda	Sessilia	*Balanus balanus*	+	+
HF4.Leray_6295260	1079	Arthropoda	Maxillopoda	Sessilia	*Balanus balanus*	+	+
HF7.Leray_1785147	2	Arthropoda	Maxillopoda	Sessilia	*Verruca stroemia*	+	-
HE1.Leray_1117391	1	Arthropoda	Pycnogonida	Pantopoda	*Endeis spinosa*	+	-
HE9.Lobo_2173983	63	Bryozoa	Gymnolaemata	Cheilostomatida	*Escharella immersa*	+	-
HF7.Leray_1838377	98	Bryozoa	Gymnolaemata	Cheilostomatida	*Membranipora membranacea*	+	-
HE3.Lobo_4881810	541	Bryozoa	Gymnolaemata	Cheilostomatida	*Scrupocellaria scruposa*	+	-
HF6.Lobo_2617384	2	Bryozoa	Gymnolaemata	Ctenostomata	*Amathia gracilis*	+	-
HF5.Lobo_3158598	5	Bryozoa	Stenolaemata	Cyclostomatida	*Crisia eburnea*	+	-
HE6.Leray_2983148	31	Chaetognatha	Sagittoidea	Aphragmophora		+	-
TS1.Leray_646185	73	Chordata	Actinopterygii	Gasterosteiformes	*Gasterosteus aculeatus*	+	+
HF4.Lobo_208606	1	Chordata	Actinopterygii	Perciformes	*Ammodytes marinus*	+	-
HF1.Leray_2487062	288	Chordata	Actinopterygii	Perciformes	*Ctenolabrus rupestris*	+	-
HF3.Lobo_3538759	472	Chordata	Actinopterygii	Perciformes	*Gobius niger*	+	-
TF1.Lobo_2807051	486	Chordata	Actinopterygii	Perciformes	*Lesueurigobius friesii*	+	+
HF9.Lobo_7596943	8	Chordata	Actinopterygii	Perciformes	*Mullus surmuletus*	+	-
HF5.Lobo_3273051	43	Chordata	Actinopterygii	Perciformes	*Trachinus draco*	+	-
HE2.Lobo_1914646	81	Chordata	Actinopterygii	Pleuronectiformes	*Limanda limanda*	+	-
HE8.Lobo_879846	265	Chordata	Actinopterygii	Pleuronectiformes	*Solea solea*	+	-
HE8.Lobo_756051	34	Chordata	Actinopterygii	Salmoniformes	*Salmo trutta*	+	-
HF3.Lobo_3595218	14	Chordata	Ascidiacea	Phlebobranchia	*Phallusia ingeria*	+	-

HE8.Lobo_873511	131011	Chordata	Leptocardii	-	*Branchiostoma lanceolatum*	+	+
TF3.Leray_6588680	3869	Cnidaria	Anthozoa	Pennatulacea	*Funiculina* sp.	+	+
TF6.Lobo_5251371	1	Cnidaria	Hydrozoa	Anthoathecata	*Corymorpha nutans*	-	+
HE9.Lobo_2164485	2	Cnidaria	Hydrozoa	Anthoathecata	*Lizzia blondina*	+	-
TF6.Leray_5512978	1481	Cnidaria	Hydrozoa	Leptothecata	*Eutima gracilis*	+	+
HF5.Lobo_3253786	232	Cnidaria	Scyphozoa	Semaeostomeae	*Aurelia aurita*	+	+
HE3.Leray_361248	14	Cnidaria	Scyphozoa	Semaeostomeae	*Cyanea capillata*	+	+
HE2.Leray_6553538	1	Cnidaria	Staurozoa	Stauromedusae		+	-
HE2.Leray_6571642	184	Cnidaria	Staurozoa	Stauromedusae	*Craterolophus convolvulus*	+	-
HE7.Leray_3802459	570	Echinodermata	Asteroidea	Forcipulatida	*Asterias rubens*	+	-
HE3.Leray_388102	85	Echinodermata	Asteroidea	Forcipulatida	*Marthasterias glacialis*	+	-
HF4.Leray_6293728	71	Echinodermata	Echinoidea	Clypeasteroida	*Echinocyamus pusillus*	+	+
HE8.Leray_7326980	315	Echinodermata	Echinoidea	Echinoida	*Psammechinus miliaris*	+	-
HE6.Lobo_7886165	1	Echinodermata	Echinoidea	Spatangoida		+	-
TF3.Leray_6591339	2079	Echinodermata	Echinoidea	Spatangoida	*Brissopsis lyrifera*	+	+
HF7.Leray_1843674	94	Echinodermata	Echinoidea	Spatangoida	*Echinocardium cordatum*	+	-
TS5.Lobo_6025603	11	Echinodermata	Holothuroidea	Dendrochirotida	*Thyone fusus*	+	+
TS3.Leray_6733304	1027065	Echinodermata	Ophiuroidea	Ophiurida		+	+
TS1.Leray_663710	3	Echinodermata	Ophiuroidea	Ophiurida	*Acrocnida brachiata*	-	+
TF1.Lobo_2726978	298	Echinodermata	Ophiuroidea	Ophiurida	*Ophiothrix fragilis*	-	+
TF1.Leray_2426830	16603	Echinodermata	Ophiuroidea	Ophiurida	*Ophiura albida*	+	+
TF5.Leray_2879711	1	Echinodermata	Ophiuroidea	Ophiurida	*Ophiura sarsii*	-	+
HF3.Leray_7012508	44	Gastrotricha	_	Macrodasyida	*Macrodasys* sp.	+	-
HE1.Lobo_948618	14	Gnathostomulida		Bursovaginoidea	*Gnathostomula armata*	+	-
TS2.Leray_4506244	1	Mollusca	Bivalvia	Lucinoida	*Thyasira equalis*	-	+

HF3.Leray_7058438	371	Mollusca	Bivalvia	Myoida	*Corbula gibba*	+	+
HE1.Lobo_894587	22	Mollusca	Bivalvia	Mytiloida	*Mytilus edulis*	+	-
TS1.Lobo_4571224	4	Mollusca	Bivalvia	Nuculida	*Nucula nucleus*	-	+
TS3.Leray_6727248	56213	Mollusca	Bivalvia	Veneroida	*Abra nitida*	+	+
HE4.Lobo_4121128	25	Mollusca	Bivalvia	Veneroida	*Dosinia lupinus*	+	+
TF5.Leray_2915847	1911	Mollusca	Bivalvia	Veneroida	*Kurtiella bidentata*	+	+
TS6.Leray_5683559	2	Mollusca	Bivalvia	Veneroida	*Lucinoma borealis*	-	+
HF1.Leray_2592679	33	Mollusca	Bivalvia	Veneroida	*Spisula subtruncata*	+	-
HE7.Leray_3779267	14392	Mollusca	Bivalvia	Veneroida	*Tellimya ferruginosa*	+	+
HF5.Lobo_3246886	1	Mollusca	Cephalopoda	Sepiida	*Sepietta neglecta*	+	-
TS1.Lobo_4750257	2	Mollusca	Gastropoda	Cephalaspidea		-	+
TS1.Lobo_4792606	2	Mollusca	Gastropoda	Cephalaspidea		-	+
HF8.Lobo_5143779	2	Mollusca	Gastropoda	Littorinimorpha	*Euspira nitida*	+	-
HE3.Lobo_4838288	34	Mollusca	Gastropoda	Neogastropoda	*Mangelia attenuata*	+	+
HF6.Lobo_2622544	37	Mollusca	Gastropoda	Neogastropoda	*Nassarius nitidus*	+	-
HE2.Lobo_1993552	50	Mollusca	Gastropoda	Nudibranchia		+	-
HE6.Leray_2935130	2	Mollusca	Gastropoda	Nudibranchia		+	-
HF1.Leray_2520121	559	Mollusca	Gastropoda	Nudibranchia	*Favorinus branchialis*	+	-
HE2.Lobo_1978270	5	Mollusca	Gastropoda	Nudibranchia	*Onchidoris muricata*	+	-
HE2.Lobo_1939813	155	Mollusca	Gastropoda	Nudibranchia	*Polycera quadrilineata*	+	-
HE2.Lobo_1938412	10	Mollusca	Gastropoda	Nudibranchia	*Polycera quadrilineata*	+	-
HF5.Leray_3991765	847	Mollusca	Gastropoda	Pulmonata	*Microhedyle glandulifera*	+	-
HF4.Leray_6295954	2965	Mollusca	Gastropoda	Sacoglossa	*Elysia viridis*	+	+
HF5.Lobo_3167773	166	Mollusca	Gastropoda	Sorbeoconcha	*Onoba semicostata*	+	-

HE4.Lobo_4138137	2	Mollusca	Gastropoda	Sorbeoconcha	*Pusillina inconspicua*	+	-
TS1.Lobo_4644275	2	Nemertea	Anopla	_	*Cerebratulus* sp.	+	+
HE4.Lobo_4203493	3	Nemertea	Palaeonemertea	_	*Carinina ochracea*	+	-
TF1.Lobo_2662495	1	Nemertea	Palaeonemertea	_	*Hubrechtella dubia*	-	+
HF7.Lobo_5876008	353	Phoronida	_	_	*Phoronis muelleri*	+	-
HE8.Lobo_843910	13	Porifera	Demospongiae	Chondrillida	*Halisarca dujardini*	+	-
HE4.Leray_3148053	1664	Porifera	Demospongiae	Suberitida	*Halichondria panicea*	+	+
TS5.Leray_1547671	2628	Priapulida	Priapulimorpha	Priapulimorphida	*Priapulus caudatus*	+	+
HF5.Leray_3885266	5	Rotifera	Eurotatoria	Flosculariaceae	*Testudinella clypeata*	+	-
HE3.Leray_357208	2	Rotifera	Monogononta	Ploima		+	-
HF8.Lobo_5184437	1	Sipuncula	Sipunculidea	Golfingiida	*Golfingia vulgaris*	+	-
TS1.Lobo_4586276	14	Xenacoelomorpha	_	Acoela	*Archaphanostoma* sp.	-	+
TS3.Lobo_1178177	4	Xenacoelomorpha	_	Acoela	*Childia macroposthium*	-	+
HF9.Lobo_7719366	2	Xenacoelomorpha	_	Acoela	*Haplogonaria viridis*	+	-
HF9.Lobo_7734506	1	Xenacoelomorpha	_	Acoela	*Notocelis Gullmarnensis*	+	-
18Sa							
OTU ID	Nb of reads	Phylum	Class	Order	*Species*	HI	GF
TF5.SSU_460284	121639	Annelida	_	_		+	+
TS3.SSU_470635	59	Annelida	_	_		-	+
HF9.SSU_7624	12	Annelida	Clitellata	Enchytraeida	*Grania* sp.	+	-
TF5.SSU_453927	2687	Annelida	Clitellata	Haplotaxida	*Tubificoides insularis*	+	+
HF3.SSU_985477	1090	Annelida	Polychaeta	_	*Aricia* sp.	+	+
HF6.SSU_322303	10	Annelida	Polychaeta	_	*Protodriloides chaetifer*	+	-

HF4.SSU_622170	1	Annelida	Polychaeta	_	*Scalibregma inflatum*	+	-
HF9.SSU_25735	3753	Annelida	Polychaeta	_	*Trilobodrilus heideri*	+	-
TS3.SSU_480632	189	Annelida	Polychaeta	Phyllodocida	*Aphrodita* sp.	-	+
HE6.SSU_371492	49226	Annelida	Polychaeta	Phyllodocida	*Brania* sp.	+	+
HE4.SSU_913344	37252	Annelida	Polychaeta	Phyllodocida	*Glycera* sp.	+	+
HF5.SSU_997904	64	Annelida	Polychaeta	Phyllodocida	*Glycinde armigera*	+	+
TS5.SSU_870099	69	Annelida	Polychaeta	Phyllodocida	*Goniada maculata*	-	+
TF6.SSU_42415	2	Annelida	Polychaeta	Phyllodocida	*Harmothoe imbricata*	-	+
HE6.SSU_350003	5	Annelida	Polychaeta	Phyllodocida	*Myrianida* sp.	+	-
HF6.SSU_324605	2	Annelida	Polychaeta	Phyllodocida	*Nereis pelagica*	+	-
HE7.SSU_239005	67220	Annelida	Polychaeta	Phyllodocida	*Pisione remota*	+	+
HE2.SSU_637269	49	Annelida	Polychaeta	Phyllodocida	*Platynereis dumerilii*	+	-
HE8.SSU_832291	1	Annelida	Polychaeta	Phyllodocida	*Progoniada regularis*	+	-
HE8.SSU_834197	1	Annelida	Polychaeta	Sabellida	*Fabriciola liguronis*	+	-
HF2.SSU_202737	4	Annelida	Polychaeta	Sabellida	*Laeospira corallinae*	+	-
HE2.SSU_640060	3	Annelida	Polychaeta	Sabellida	*Myriochele* sp.	+	-
TS5.SSU_869292	123	Annelida	Polychaeta	Spionida	*Apistobranchus* sp.	-	+
TS3.SSU_517096	1407	Annelida	Polychaeta	Spionida	*Laonice* sp.	-	+
HE3.SSU_123438	1952	Annelida	Polychaeta	Spionida	*Spio* sp.	+	+
TS5.SSU_882766	60	Annelida	Polychaeta	Terebellida	*Diplocirrus glaucus*	-	+
HF2.SSU_193854	1	Annelida	Polychaeta	Terebellida	*Flabelligera* sp.	+	-
TF6.SSU_63146	669	Annelida	Polychaeta	Terebellida	*Pectinaria* sp.	-	+
TS5.SSU_883475	4155	Annelida	Polychaeta	Terebellida	*Terebellides stroemii*	-	+
TF4.SSU_139713	193	Arthropoda	Branchiopoda	_		-	+

HE5.SSU_184679	149	Arthropoda	Malacostraca	_		+	-
HE8.SSU_832214	1	Arthropoda	Malacostraca	Decapoda	*Nikoides* sp.	+	-
HF5.SSU_994971	7	Arthropoda	Malacostraca	Decapoda	*Praebebalia longidactyla*	+	-
TF6.SSU_56595	65992	Arthropoda	Maxillopoda	_		+	+
HF9.SSU_15855	31800	Arthropoda	Maxillopoda	_		+	+
HF2.SSU_208480	21241	Arthropoda	Maxillopoda	_		+	+
TS2.SSU_812824	433	Arthropoda	Maxillopoda	_		+	+
TF3.SSU_955499	185	Arthropoda	Maxillopoda	_		+	+
TF5.SSU_470101	360	Arthropoda	Maxillopoda	*Harpacticoida*	*Typhlamphiascus typhlops*	-	+
HE1.SSU_864375	1160	Arthropoda	Ostracoda	Podocopida	*Hemicytherura kajiyamai*	+	+
HE7.SSU_253407	2584	Arthropoda	Ostracoda	Podocopida	*Loxocorniculum mutsuense*	+	+
HE5.SSU_181011	1	Arthropoda	Pycnogonida	Pantopoda	*Anoplodactylus californicus*	+	-
HE2.SSU_646490	123	Arthropoda	Pycnogonida	Pantopoda	*Callipallene* sp.	+	-
HE2.SSU_638224	23	Bryozoa	_	_		+	-
HE6.SSU_373369	2	Bryozoa	Stenolaemata	Cyclostomatida	*Plagioecia patina*	+	-
HE1.SSU_850917	4	Bryozoa	Stenolaemata	Cyclostomatida	*Tubulipora lobifera*	+	-
TF5.SSU_412099	18	Cephalorhyncha	Kinorhyncha	Homalorhagida	*Pycnophyes kielensis*	-	+
HE7.SSU_239963	45	Chordata	Actinopteri	Perciformes	*Hypseleotris* sp.	+	+
HE3.SSU_123107	4	Chordata	Ascidiacea	_		+	-
HF9.SSU_12142	727	Chordata	Ascidiacea	Phlebobranchia	*Ascidiella* sp.	+	+
HF4.SSU_611685	114	Chordata	Ascidiacea	Phlebobranchia	*Corella inflata*	+	+
HE2.SSU_639404	209	Chordata	Ascidiacea	Stolidobranchia	*Molgula* sp.	+	-
HE9.SSU_314754	616	Chordata	Ascidiacea	Stolidobranchia	*Styela plicata*	+	-
HE8.SSU_834024	11058	Chordata	Leptocardii	_	*Branchiostoma* sp.	+	-
TF1.SSU_674740	2212	Cnidaria	Anthozoa	Actiniaria	*Nematostella vectensis*	+	+
TS3.SSU_472524	2741	Cnidaria	Hydrozoa	_		+	+

TS3.SSU_518760	7860	Cnidaria	Hydrozoa	Anthoathecata	*Euphysa* sp.	+	+
HE2.SSU_639670	1	Cnidaria	Hydrozoa	Leptothecatha	*Abietinaria filicula*	+	-
TF4.SSU_152912	61418	Echinodermata	_	_		+	+
HE5.SSU_186025	8038	Echinodermata	_	_		+	+
TF4.SSU_155631	5491	Echinodermata	_	_		+	+
TS5.SSU_881395	25	Echinodermata	_	_		-	+
HE4.SSU_914821	1	Echinodermata	Holothuroidea	Apodida	*Leptosynapta* sp.	+	-
HF9.SSU_2577	1006	Gastrotricha	_	Chaetonotida	*Chaetonotus* sp.	+	+
HE7.SSU_244283	249	Gastrotricha	_	Macrodasyida	*Diplodasys meloriae*	+	-
HF5.SSU_996540	161	Gastrotricha	_	Macrodasyida	*Lepidodasys* sp.	+	-
HF5.SSU_995416	636	Gastrotricha	_	Macrodasyida	*Macrodasys* sp.	+	-
HF2.SSU_192734	479	Gastrotricha	_	Macrodasyida	*Macrodasys* sp.	+	-
HF7.SSU_385728	6934	Gastrotricha	_	Macrodasyida	*Mesodasys* sp.	+	+
HE7.SSU_242889	3013	Gastrotricha	_	Macrodasyida	*Tetranchyroderma thysanophorum*	+	-
HF1.SSU_770513	339	Gastrotricha	_	Macrodasyida	*Thaumastoderma ramuliferum*	+	-
HF1.SSU_760431	5	Gastrotricha	_	Macrodasyida	*Urodasys* sp.	+	-
TF6.SSU_44832	3816	Mollusca	Bivalvia	_		+	+
HF2.SSU_208561	14	Mollusca	Bivalvia	Anomalodesmata		+	+
HF8.SSU_788507	1	Mollusca	Bivalvia	Limoida	*Limaria hians*	+	-
TF3.SSU_924397	11725	Mollusca	Bivalvia	Veneroida	*Abra* sp.	+	+
HE9.SSU_317977	1982	Mollusca	Bivalvia	Verenoida	*Arctica islandica*	+	+
TF4.SSU_132537	1581	Mollusca	Gastropoda	Neogastropoda	*Nassarius festivus*	+	+
HF1.SSU_779114	65	Nematoda	Chromadorea	Araeolaimida	*Odontophora* sp.	+	+
TF6.SSU_48167	2940	Nematoda	Chromadorea	Araeolaimida	*Sabatieria* sp.	+	+
TF1.SSU_710679	639	Nematoda	Chromadorea	Chromadorida		+	+
HF2.SSU_192072	2	Nematoda	Chromadorea	Chromadorida	*Chromadora nudicapitata*	+	-
HF1.SSU_759758	4	Nematoda	Chromadorea	Plectida		+	-
HF9.SSU_20251	636	Nematoda	Desmodorida	Microlaimidae		+	+

HE3.SSU_124287	13	Nematoda	Enoplea	Enoplida	*Enoploides* sp.	+	-
HE3.SSU_110275	8	Nematoda	Enoplea	Enoplida	*Enoplus* sp.	+	-
HE5.SSU_188855	27	Nematoda	Enoplea	Enoplida	*Symplocostoma* sp.	+	+
TS6.SSU_587229	493	Nematoda	Enoplea	Enoplida	*Viscosia viscosa*	+	+
TF3.SSU_938615	642	Nemertea	_	_		+	+
TF6.SSU_49192	265	Nemertea	Anopla	_	*Cerebratulus marginatus*	+	+
HE4.SSU_908113	877	Nemertea	Anopla	_	*Lineus bilineatus*	+	+
HF9.SSU_3582	6	Nemertea	Paleonemertea	_	*Callinera grandis*	+	-
HE3.SSU_121696	12053	Nemertea	Paleonemertea	_	*Cephalothrix filiformis*	+	+
TF5.SSU_434928	1760	Nemertea	Paleonemertea	_	*Hubrechtella dubia*	+	+
TS2.SSU_818002	1	Platyhelminthes	Rhabditophora	Cestoda		-	+
HE9.SSU_303121	1939	Platyhelminthes	Rhabditophora	Haplopharyngida	*Haplopharynx rostratus*	+	-
HF1.SSU_773830	1	Platyhelminthes	Rhabditophora	Prolecithophora	*Allostoma neostiliferum*	+	-
HE2.SSU_650311	8	Platyhelminthes	Rhabditophora	Prolecithophora	*Cylindrostoma* sp.	+	-
HE5.SSU_177399	4	Platyhelminthes	Rhabditophora	Prolecithophora	*Euxinia baltica*	+	-
HF9.SSU_23023	8367	Platyhelminthes	Rhabditophora	Prolecithophora	*Plagiostomum cinctum*	+	+
TS2.SSU_822141	938	Platyhelminthes	Rhabditophora	Prolecithophora	*Plagiostomum cuticulata*	-	+
TF6.SSU_52738	214	Platyhelminthes	Rhabditophora	Prolecithophora	*Plagiostomum striatum*	-	+
TF5.SSU_433159	2	Platyhelminthes	Rhabditophora	Prolecithophora	*Ulianinia mollissima*	-	+
HF9.SSU_24513	59	Platyhelminthes	Rhabditophora	Proseriata	*Monocelis lineata*	+	+
HF2.SSU_201740	2	Platyhelminthes	Rhabditophora	Rhabdocoela	*Phonorhynchus helgolandicus*	+	-
TS6.SSU_592673	245	Platyhelminthes	Rhabditophora	Rhabdocoela	*Proxenetes* sp.	+	+
HF4.SSU_616041	771	Platyhelminthes	Rhabditophora	Seriata		+	-
HE3.SSU_117223	181	Porifera	Calcarea	_		+	+

HE7.SSU_223989	12	Porifera	Demospongiae	Chondrillida	*Halisarca dujardini*	+	-
HF9.SSU_26977	8	Porifera	Demospongiae	Clionaida	*Spheciospongia vesparium*	+	-
HE6.SSU_383060	3	Sipuncula	Sipunculidea	Golfingiida	*Phascolopsis gouldii*	+	-
HE6.SSU_348954	2	Tardigrada	Eutardigrada	Parachela	*Halobiotus crispae*	+	-
TF3.SSU_927927	2	Xenacoelomorpha	_	_		-	+
HE3.SSU_116025	28	Xenacoelomorpha	_	Acoela	*Archaphanostoma* sp.	+	+
HF9.SSU_26335	1	Xenacoelomorpha	_	Acoela	*Archaphanostoma* sp.	+	-
TS2.SSU_815721	2	Xenacoelomorpha	_	Acoela	*Childia* sp.	-	+
TS2.SSU_815970	1	Xenacoelomorpha	_	Acoela	*Childia* sp.	-	+
HF2.SSU_190395	2386	Xenacoelomorpha	_	Acoela	*Eumecynostomum* sp.	+	-
HF1.SSU_758202	74	Xenacoelomorpha	_	Acoela	*Haplogonaria* sp.	+	+
HF9.SSU_13290	5	Xenacoelomorpha	_	Nemertodermatida	*Flagellophora apelti*	+	-
TS6.SSU_601153	28	Xenacoelomorpha	_	Nemertodermatida	*Nemertoderma westbladi*	-	+

Invasive and alien species detected in the samples

Five alien species were detected in in the sample, of which two are considered invasive (in bold; Table 9), and the other three are on alert lists. The two invasive species (*Acartia tonsa*, a copepod, and *Alexandrium ostenfefeldii*, a dinoflagellate) could easily be overlooked in routine monitoring programs. Species within the genus *Acartia* are difficult to distinguish (Jensen 2010) and the invasive species can be confused with other native species. Also *A. ostenfeldii* is easily misidentified as other *Alexandrium* species; detailed thecal plate observation is often necessary for proper identification (Balech 1995). This shows the potential of molecular techniques for monitoring invasive species, and points to problems using traditional identification techniques. Many invasive species arrive in an area as spores, larvae or juveniles - all life stages that may be easily overlooked and problematic to identify to species level. Target barcoding of environmental DNA (eDNA) shows a great promise for detecting species without the need of costly sampling schemes. This would also allow for more random sampling in an area, increasing the probability of actually finding a species even when they occur in low numbers.

Table 9.

Invasive species (in bold) and species on alert lists (not bold) found in the samples. X indicates where the species were found.

Species	**Phylum**	**COI**		**18S**	
		Hållö island	Gullmarn Fjord	Hållö island	Gullmarn Fjord
Acartia tonsa	Arthropoda	x	x		
Alexandrium ostenfeldii	Dinoflagellata			x	x
Bonnemaisonia hamifera	Rhodophyta	x	x	x	
Penilia avirostris	Arthropoda	x	x		
Thalassiosira punctigera	Bacillariophyta	x			

Comparison of metabarcoding versus morphology-based identification of Xenacoelomorpha

Comparison of morphology-based assessment of Xenacoelomorpha diversity with metabarcoding using taxonomic assignments to the phylum level (with 80% similarity threshold; Suppl. materials 2, 3), shows that extraction procedures have strong impact on the effectiveness of morphology-based identification (Tables 10, 11). Using freshwater for extraction of Xenacoelomorpha rendered most of them unrecognizable and unidentifiable, but left their DNA intact and suitable for metabarcoding. No identifiable Xenacoelomorpha were found in the Hållö samples extracted using flotation with fresh water, while all specimens found in Gullmarn Fjord were treated together as one taxon "*Acoela* sp." for the lack of better alternative. Metabarcoding, on the other hand, recovered between 6 and 15 taxa (OTUs) from the Hållö samples extracted using flotation with fresh water (Table 11), and up to 13 taxa (OTUs) from the same type of samples from the Gullmarn Fjord site (Table 11), depending on the barcoding region used. Just like for nematodes (see below), 18S barcodes always gave higher overall estimates of diversity (number of OTUs) compared to COI (Table 11). 18S also gave higher diversity estimates, compared to morphology-based identification for the Hållö samples extracted using flotation with MgCl2 (11 versus 7), but lower for the Gullmarn Fjord site samples extracted using siphoning (9 versus 15). COI Leray primers were less effective compared to the COI Lobo primers that recovered 2-6 OTUs more in all samples (Table 11). The most numerous of the morphologically identified species, *Mecynostomum tenuissimum*, was present with 120 specimens in the manually sorted samples, but was not detected at all in the 18S samples. Note that the 18S and COI sequences for all of the species identified in the visually sorted samples are present in the reference database. This raises the question of the efficiacy of using the SSU_FO4-SSU_R22 18 S fragment for metabarcoding of acoelomorphs. A recent study found a number of unknown xenacoelomorph taxa while data mining metabarcoding sequences from surveys of pelagial and deep benthic habitats (Arroyo et al. 2016). Unknown xenacoelomorph species may exist also at the moderate sampling depths we sampled in the Gullmarn Fjord. Our siphoning technique relies on migration of specimens to the sediment surface in response to hypoxia. It is possible that

there are xenacoelomorphs with high tolerance for hypoxia that are not captured by the siphoning method, and thus would not be found in the manually sorted samples, but could be detected by metabarcoding of unprocessed samples. It should be noted that the extraction method used on the Hållö samples does not rely on migration of specimens to the surface.

Table 10.

Taxonomic composition and relative abundance (% of the total number of specimens) of Xenacoelomorpha species in Gullmarn Fjord and Hållö sites.

		Gullmarn Fjord		Hållö	
	Taxon	**Siphoning**	**Flotation with fresh water**	**Flotation with MgCl2 solution**	**Flotation with fresh water**
	Acoela				
1	*Haploposthia rubropunctata*	1.03	0	0	0
2	*Childia brachyposthium*	3.78	0	0	0
3	*Childia submaculatum*	1.03	0	0	0
4	*Childia trianguliferum*	2.06	0	0	0
5	*Childia crassum*	3.44	0	0	0
6	*Childia* sp.	25.09	0	0	0
7	*Mecynostomum tenuissimum*	43.99	0	0	0
8	*Mecynostomum auritum*	0.34	0	0	0
9	cf. *Eumecynostomum altitudi*	4.81	0	0	0
10	*Philactinoposthia* sp.	0.34	0	0	0
11	Acoela sp.	2.06	100	88.71	0
12	*Faerlea glomerata*	3.09	0		
13	*Archaphanostoma* sp.	0.34	0	0.81	0
14	*Postmecynostomum glandulosum*	0	0	2.42	0
15	*Paramecynostomum* sp.	0	0	0.81	0
16	*Eumecynostomum macrobursalium*	0	0	0.81	0
17	*Isodiametra* sp.	0	0	0.81	0
18	*Haplogonaria viridis/Archocelis macrorhabditis*	0	0	5.65	0
	Nemertodermatida				
19	*Nemertoderma westbladi*	8.25	0	0	0
20	*Flagellophora apelti*	0.34	0	0	0

Table 11.

Total number of Xenacoelomorpha taxa or OTUs distinguished based on morphology (Table 10), 18S and COI from different sampling sites and extraction methods (placement of OTUs is based on 80% similarity threshold, Suppl. materials 2, 3)

Site / extraction method	morphology-based	18S	COI (Lobo)	COI (Leray)
Hållo, flotation with MgCl2	7	11	8	6
Hållö, flotation with fresh water	0	15	11	6
Hållö, total	7	16	12	7
Gullmarn Fjord, siphoning	15	11	9	4
Gullmarn Fjord, flotation with fresh water	1	13	2	0
Gullmarn Fjord, total	15	19	10	4

Comparison of metabarcoding versus morphology-based identification of Nematoda

Both study sites are characterized by rich and diverse nematode fauna. The Hållö site had a total of 107 species of nematodes, belonging to 86 genera (Holovachov et al. 2017). Of these, 88 species belonging to 73 genera were found in samples extracted by flotation with a MgCl2 solution, and 101 species belonging to 83 genera were found in samples extracted by flotation with fresh water. The Gullmarn fjord site had a total of 113 nematode species of nematodes, belonging to 77 genera (Holovachov et al. 2017). Of these, 81 species belonging to 62 genera were found in samples extracted by siphoning, and 102 species belonging to 70 genera were found in samples extracted by flotation with fresh water. A certain small number of nematode individuals in each sample were not identified to species/genus/family, either due to their developmental stage or quality of preservation.

The final list of nematode OTUs includes 139 18S sequences. Only two 18S OTUs were positively identified using QIIME to species level using 97% similarity threshold: *Viscosia viscosa* (TS6.SSU58722) and *Chromadora nudicapitata* (HF2.SSU192072), six more were assigned to reference sequences identified to genus level only (Suppl. material 1). Only 22 COI sequences were assigned to the phylum Nematoda, and none was identified to species level.

When comparing the results of morphology-based assessment of nematode diversity with metabarcoding using taxonomic assignments to the phylum level in this particular study (with 80% similarity threshold; Suppl. materials 2, 3), the detailed and extensive examination of samples and morphology-based species identification provided more comprehensive estimates of nematode diversity (107 species in Hållö and 113 species in Gullmarn Fjord) than metabarcoding using either one of the molecular markers, independently of the extraction technique or locality (Table 12). Moreover, COI barcodes were much harder to obtain for marine nematodes using either one of the primers (16 OTUs in Hållö and 9 OTUs in Gullmarn Fjord using Lobo primers; 17 OTUs in Hållö and 4 OTUs in Gullmarn Fjord using Leray primers), comparing to 18S (95 OTUs in Hållö and 78

OTUs in Gullmarn Fjord site; Table 12). Due to the very limited reference databases available for marine nematodes, very few nematode OTUs can be identified to species or genus level, making it difficult to use metabarcoding data in ecological studies.

Table 12.

Total number of nematode taxa or OTUs distinguished based on morphology (after Holovachov et al. 2017), 18S and COI from different sampling sites and extraction methods (placement of OTUs is based on 80% similarity threshold, Suppl. materials 2, 3)

Site / extraction method	morphology-based	18S	COI (Lobo)	COI (Leray)
Hållo, flotation with MgCl2	88	71	12	11
Hållö, flotation with fresh water	101	78	14	14
Hållö, total	107	95	16	17
Gullmarn Fjord, siphoning	81	47	8	4
Gullmarn Fjord, flotation with fresh water	102	67	4	2
Gullmarn Fjord, total	113	78	9	4

Acknowledgements

We would like to thank the Genomics Core facility platform at the Sahlgrenska Academy, University of Gothenburg. The SweBoL (Swedish Barcode of Life) network and Christer Erséus are thanked for sharing barcode databases of Swedish invertebrates. We would also like to thank Nicolas Girard for help with scripting. This work was in part supported by the project "Systematics of Swedish free-living nematodes of the orders Desmodorida and Araeolaimida" (Swedish Taxonomy Initiative, ArtDatabanken, Sweden) awarded to OH, and by the Swedish Research Council project (2012-3446) 'Biodiversity genomics: Species identification pipelines for analyzing marine invertebrate larval stages, community structure, and trophic interactions' awarded to SJB.

References

- Arroyo A, López-Escardó D, Vargas Cd, Ruiz-Trillo I (2016) Hidden diversity of Acoelomorpha revealed through metabarcoding. Biology Letters 12 (9): 20160674. https://doi.org/10.1098/rsbl.2016.0674
- Balech E (1995) The Genus *Alexandrium* Halim (Dinoflagellata). Sherkin Island Marine Station, Ireland ISBN: 1 870492 61 7.
- Bik HM, Halanych KM, Sharma J, Thomas WK (2012a) Dramatic Shifts in Benthic Microbial Eukaryote Communities following the Deepwater Horizon Oil Spill. Plos One 7 (6): e38550. [In English]. https://doi.org/10.1371/journal.pone.0038550
- Bik HM, Sung W, De Ley P, Baldwin JG, Sharma J, Rocha-Olivares A, Thomas WK (2012b) Metagenetic community analysis of microbial eukaryotes illuminates

biogeographic patterns in deep-sea and shallow water sediments. Molecular Ecology 21 (5): 1048-1059. [In English]. https://doi.org/10.1111/j.1365-294X.2011.05297.x

- Bohmann K, Evans A, Gilbert MTP, Carvalho GR, Creer S, Knapp M, Yu DW, de Bruyn M (2014) Environmental DNA for wildlife biology and biodiversity monitoring. Trends in ecology & evolution 29 (6): 358-67. https://doi.org/10.1016/j.tree.2014.04.003
- Bourlat SJ, Haenel Q, Finnman J, Leray M (2016) Metabarcoding of marine eukaryotes: Illumina Mi-Seq library preparation using fusion primer methods. In: Bourlat SJ (Ed.) Marine Genomics - Methods and protocols. Springer, New York. [ISBN 978-1-4939-3772-1].
- Brannock PM, Halanych KM (2015) Meiofaunal community analysis by high-throughput sequencing: Comparison of extraction, quality filtering, and clustering methods. Marine Genomics In Press.
- Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Pena AG, Goodrich JK, Gordon JI, Huttley GA, Kelley ST, Knights D, Koenig JE, Ley RE, Lozupone CA, McDonald D, Muegge BD, Pirrung M, Reeder J, Sevinsky JR, Turnbaugh PJ, Walters WA, Widmann J, Yatsunenko T, Zaneveld J, Knight R (2010) QIIME allows analysis of high-throughput community sequencing data. Nat Methods 7 (5): 335-6. [In eng]. https://doi.org/10.1038/nmeth.f.303
- Cowart DA, Pinheiro M, Mouchel O, Maguer M, Grall J, Mine J, Arnaud-Haond S (2015) Metabarcoding Is Powerful yet Still Blind: A Comparative Analysis of Morphological and Molecular Surveys of Seagrass Communities. Plos One 10 (2): e0117562. [In English]. https://doi.org/10.1371/Journal.Pone.0117562
- Creer S, Fonseca VG, Porazinska DL, Giblin-Davis RM, Sung W, Power DM, Packer M, Carvalho GR, Blaxter ML, Lambshead PJD, Thomas WK (2010) Ultrasequencing of the meiofaunal biosphere: practice, pitfalls and promises. Molecular Ecology 19: 4-20. [In English]. https://doi.org/10.1111/J.1365-294x.2009.04473.X
- De Grisse AT (1969) Redescription ou modifications de quelques techniques utilisées dans l'etude des nematodes phytoparasitaires. Mededelingen Rijksfakulteit Landbouwwetenschappen Gent 34: 351-369.
- de Vargas C, Audic S, Henry N, Decelle J, Mahe F, Logares R, Lara E, Berney C, Le Bescot N, Probert I, Carmichael M, Poulain J, Romac S, Colin S, Aury JM, Bittner L, Chaffron S, Dunthorn M, Engelen S, Flegontova O, Guidi L, Horak A, Jaillon O, Lima-Mendez G, Lukes J, Malviya S, Morard R, Mulot M, Scalco E, Siano R, Vincent F, Zingone A, Dimier C, Picheral M, Searson S, Kandels-Lewis S, Acinas SG, Bork P, Bowler C, Gorsky G, Grimsley N, Hingamp P, Iudicone D, Not F, Ogata H, Pesant S, Raes J, Sieracki ME, Speich S, Stemmann L, Sunagawa S, Weissenbach J, Wincker P, Karsenti E, Coordinators TO (2015) Eukaryotic plankton diversity in the sunlit ocean. Science 348 (6237): . [In English]. https://doi.org/10.1126/Science.1261605
- Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26 (19): 2460-1. https://doi.org/10.1093/bioinformatics/btq461
- Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R (2011) UCHIME improves sensitivity and speed of chimera detection. Bioinformatics 27 (16): 2194-200. https://doi.org/10.1093/bioinformatics/btr381
- Fadrosh DW, Ma B, Gajer P, Sengamalay N, Ott S, Brotman RM, Ravel J (2014) An improved dual-indexing approach for multiplexed 16S rRNA gene sequencing on the Illumina MiSeq platform. Microbiome 2 (1): . https://doi.org/10.1186/2049-2618-2-6

- Folmer O, Black M, Hoeh W, Lutz R, Vrijenhoek R (1994) DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. Mol Mar Biol Biotechnol 3 (5): 294-9. [In eng]. URL: http://www.ncbi.nlm.nih.gov/pubmed/7881515
- Fonseca VG, Carvalho GR, Nichols B, Quince C, Johnson HF, Neill SP, Lambshead JD, Thomas WK, Power DM, Creer S (2014) Metagenetic analysis of patterns of distribution and diversity of marine meiobenthic eukaryotes. Global Ecology and Biogeography 23 (11): 1293-1302. [In English]. https://doi.org/10.1111/geb.12223
- Fonseca VG, Carvalho GR, Sung W, Johnson HF, Power DM, Neill SP, Packer M, Blaxter ML, Lambshead PJD, Thomas WK, Creer S (2010) Second-generation environmental sequencing unmasks marine metazoan biodiversity. Nature Communications 1[In English]. https://doi.org/10.1038/Ncomms1095
- Giongo A, Davis-Richardson A, Crabb D, Triplett E (2010) TaxCollector: Modifying Current 16S rRNA Databases for the Rapid Classification at Six Taxonomic Levels. Diversity 2 (7): 1015-1025. https://doi.org/10.3390/d2071015
- Hao XL, Jiang R, Chen T (2011) Clustering 16S rRNA for OTU prediction: a method of unsupervised Bayesian clustering. Bioinformatics 27 (5): 611-618. [In English]. https://doi.org/10.1093/Bioinformatics/Btq725
- Hebert PDN, Ratnasingham S, deWaard JR (2003) Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. Proceedings of the Royal Society of London Series B-Biological Sciences 270[In English]. https://doi.org/10.1098/Rsbl.2003.0025
- Higgins RPTH (1988) Introduction to the study of Meiofauna. Smithsoniam Institution Press, Washington D.C., London
- Holovachov O, Haenel Q, Bourlat SJ, Jondelius U (2017) The choice of taxonomy assignment approach has strong impact on the efficiency of identification of anonymous metabarcodes of marine nematodes. Manuscript in preparation.
- Jensen K (2010) NOBANIS – Invasive Alien Species Fact Shee*t: Acartia tonsa* From: Identification key to marine invasive species in Nordic waters. www.nobanis.org. Accessed on: 2017-3-01.
- Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C, Thierer T, Ashton B, Meintjes P, Drummond A (2012) Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 28 (12): 1647-9. https://doi.org/10.1093/bioinformatics/bts199
- Kircher M, Sawyer S, Meyer M (2012) Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. Nucleic Acids Res 40 (1): . https://doi.org/10.1093/nar/gkr771
- Lallias D, Hiddink JG, Fonseca VG, Gaspar JM, Sung W, Neill SP, Barnes N, Ferrero T, Hall N, Lambshead PJD, Packer M, Thomas WK, Creer S (2015) Environmental metabarcoding reveals heterogeneous drivers of microbial eukaryote diversity in contrasting estuarine ecosystems. Isme Journal 9 (5): 1208-1221. [In English]. https://doi.org/10.1038/ismej.2014.213
- Lanzén A, Jørgensen S, Huson D, Gorfer M, Grindhaug SH, Jonassen I, Øvreås L, Urich T (2012) CREST – Classification Resources for Environmental Sequence Tags. PLoS ONE 7 (11): e49334. https://doi.org/10.1371/journal.pone.0049334

- Lawson Handley L (2015) How will the 'molecular revolution' contribute to biological recording? Biological Journal of the Linnean Society 115 (3): 750-766. https://doi.org/10.1111/bij.12516
- Leray M, Knowlton N (2015) DNA barcoding and metabarcoding of standardized samples reveal patterns of marine benthic diversity. Proceedings of the National Academy of Sciences of the United States of America 112 (7): 2076-2081. [In English]. https://doi.org/10.1073/pnas.1424997112
- Leray M, Yang JY, Meyer CP, Mills SC, Agudelo N, Ranwez V, Boehm JT, Machida RJ (2013) A new versatile primer set targeting a short fragment of the mitochondrial COI region for metabarcoding metazoan diversity: application for characterizing coral reef fish gut contents. Frontiers in Zoology 10[In English]. https://doi.org/10.1186/1742-9994-10-34
- Lindeque PK, Parry HE, Harmer RA, Somerfield PJ, Atkinson A (2013) Next Generation Sequencing Reveals the Hidden Diversity of Zooplankton Assemblages. Plos One 8 (11): e81327. [In English]. https://doi.org/10.1371/journal.pone.0081327
- Lobo J, Costa PM, Teixeira MA, Ferreira MS, Costa MH, Costa FO (2013) Enhanced primers for amplification of DNA barcodes from a broad range of marine metazoans. BMC Ecol 13 https://doi.org/10.1186/1472-6785-13-34
- Moreno M, Vezzulli L, Marin V, Laconi P, Albertelli G, Fabiano M (2008) The use of meiofauna diversity as an indicator of pollution in harbours. Ices Journal of Marine Science 65 (8): 1428-1435. [In English]. https://doi.org/10.1093/icesjms/fsn116
- Porazinska DL, Giblin-Davis RM, Faller L, Farmerie W, Kanzaki N, Morris K, Powers TO, Tucker AE, Sung W, Thomas WK (2009) Evaluating high-throughput sequencing as a method for metagenomic analysis of nematode diversity. Mol Ecol Resour 9 (6): 1439-50. https://doi.org/10.1111/j.1755-0998.2009.02611.x
- Shokralla S, Spall JL, Gibson JF, Hajibabaei M (2012) Next-generation sequencing technologies for environmental DNA research. Molecular Ecology 21 (8): 1794-1805. [In English]. https://doi.org/10.1111/J.1365-294x.2012.05538.X
- Vincx M (1996) Meiofauna in marine and freshwater sediments. In: Hall GS (Ed.) Methods for the examination of organismal diversity in soils and sediments.
- Zhang H, Gao S, Lercher MJ, Hu S, Chen WH (2012) EvolView, an online tool for visualizing, annotating and managing phylogenetic trees. Nucleic Acids Res 40: 569-72. https://doi.org/10.1093/nar/gks576

Supplementary materials

Suppl. material 1: OTUs identified to species level in the samples using 97% sequence similarity, all organism groups

Authors: Quiterie Haenel, Oleksandr Holovachov, Ulf Jondelius, Per Sundberg and Sarah J. Bourlat

Data type: Occurrence records from Metabarcoding for Hållö island and Gullmarsfjord, Sweden.

Brief description: Sequence similarity search at 97% similarity allowed us to identify some OTUs to species level. 215 COI OTUs and 243 18S OTUs were identified to species from both sites (Hållö island and Gullmarsfjord).

Filename: TableS1.xlsx - Download file (85.61 kb)

Suppl. material 2: OTU table for 18S

Authors: Quiterie Haenel, Oleksandr Holovachov, Ulf Jondelius, Per Sundberg and Sarah J. Bourlat
Data type: Metagenomic, OTU table
Brief description: OTU table showing all 18S OTUs, their taxonomic assignment at 80% similarity and number of reads per sample (HE: Hållö Flotation, HF: Hållö Flotation MgCl2, TS: Gullmarn Fjord Siphoning, TF: Gullmarn Fjord Flotation)
Filename: 18S_otu_table.txt - Download file (318.45 kb)

Suppl. material 3: OTU table for COI

Authors: Quiterie Haenel, Oleksandr Holovachov, Ulf Jondelius, Per Sundberg and Sarah J. Bourlat
Data type: Metagenomic, OTU table
Brief description: OTU table showing all COI OTUs, their taxonomic assignment at 80% similarity and number of reads per sample (HE: Hållö Flotation, HF: Hållö Flotation MgCl2, TS: Gullmarn Fjord Siphoning, TF: Gullmarn Fjord Flotation)
Filename: CO1_otu_table.txt - Download file (728.63 kb)

Chapter 6

The choice of taxonomy assignment approach has a strong impact on the efficiency of the identification of anonymous metabarcodes of marine nematodes

Holovachov et al. 2017, Royal Society Open Science

ROYAL SOCIETY OPEN SCIENCE

rsos.royalsocietypublishing.org

Research

Cite this article: Holovachov O, Haenel Q, Bourlat SJ, Jondelius U. 2017 Taxonomy assignment approach determines the efficiency of identification of OTUs in marine nematodes. *R. Soc. open sci.* **4**: 170315. http://dx.doi.org/10.1098/rsos.170315

Received: 6 April 2017
Accepted: 18 July 2017

Subject Category:
Biology (whole organism)

Subject Areas:
taxonomy and systematics/ecology

Keywords:
biodiversity, identification, barcode, nematodes, metabarcoding, meiobenthos

Author for correspondence:
Oleksandr Holovachov
e-mail: oleksandr.holovachov@nrm.se

Electronic supplementary material is available online at https://doi.org/10.6084/m9.figshare.c.3844801.

THE ROYAL SOCIETY PUBLISHING

Taxonomy assignment approach determines the efficiency of identification of OTUs in marine nematodes

Oleksandr Holovachov[1], Quiterie Haenel[2], Sarah J. Bourlat[3] and Ulf Jondelius[1]

[1]Department of Zoology, Swedish Museum of Natural History, Stockholm, Sweden
[2]Zoological Institute, University of Basel, Basel, Switzerland
[3]Department of Marine Sciences, University of Gothenburg, Gothenburg, Sweden

OH, 0000-0002-4285-0754

Precision and reliability of barcode-based biodiversity assessment can be affected at several steps during acquisition and analysis of data. Identification of operational taxonomic units (OTUs) is one of the crucial steps in the process and can be accomplished using several different approaches, namely, alignment-based, probabilistic, tree-based and phylogeny-based. The number of identified sequences in the reference databases affects the precision of identification. This paper compares the identification of marine nematode OTUs using alignment-based, tree-based and phylogeny-based approaches. Because the nematode reference dataset is limited in its taxonomic scope, OTUs can only be assigned to higher taxonomic categories, families. The phylogeny-based approach using the evolutionary placement algorithm provided the largest number of positively assigned OTUs and was least affected by erroneous sequences and limitations of reference data, compared to alignment-based and tree-based approaches.

1. Introduction

Metabarcoding studies based on high-throughput sequencing of amplicons from marine samples have reshaped our understanding of the biodiversity of marine microscopic eukaryotes, revealing a much higher diversity than previously known [1]. Early metabarcoding of the slightly larger sediment-dwelling meiofauna has mainly focused on scoring the relative diversity of taxonomic groups [1–3]. The next step in metabarcoding, identification of species, is limited by the available reference database, which is sparse for most marine taxa, and by the matching algorithms.

In this paper, we are evaluating to what extent sequences of unidentified putative species (operational taxonomic units, OTUs) of marine nematodes can be assigned to family-level taxa using publicly available reference sequences, and which of three matching strategies, alignment-based, tree-based or phylogeny-based, provides the highest number of identified OTUs.

The reference datasets for marine nematodes are sparsely populated, as correctly pointed out in Dell'Anno *et al*. [4]. The most recent check of NCBI GenBank (February 2017) reveals that less than 180 genera and about 170 identified species of marine nematodes are included, compared to over 530 described genera and almost 4750 described species (based on [5] with updates). This summarized number of records in GenBank does not take into consideration which genes are represented (mostly near complete or partial 18S and partial 28S rDNA), but gives the total number of entries. Not all of these entries include sequences suitable to be used as references for metabarcoding. As completeness of the reference databases for marine nematodes is insufficient to assign all OTUs to species level [6], one has to consider if they can be assigned to taxonomic categories above species level, and if this type of data can be used in research.

Assignment of OTUs to nematode genera faces the same problem as the assignment of OTUs to species—limited representation of identified taxa in reference databases (see above). Identification to the family level of those OTUs that cannot be assigned to any particular species or genus is the next best option. It provides enough information to group nematode OTUs into trophic [7,8] and functional [9] groups and apply ecological metrics, such as Maturity Index [10], used to evaluate the complexity and functioning of nematode communities [11]. This approach has already been applied in metabarcoding studies of terrestrial nematode communities from the Arctic and the tropics [12,13].

Although it would be possible to generate new barcodes for marine nematodes from our study sites to supplement existing reference datasets, the purpose of this paper is to follow the typical scenario when metabarcoding projects rely on existing databases and do not publish new reference sequences.

Identification of OTUs can be done using a number of currently available approaches and applications, several of which will be tested and compared below. In general, all taxonomy assignment methods can be grouped into four categories: alignment-based, probabilistic, tree-based and phylogeny-based.

Alignment-based approaches use various measures of similarity between query and reference sequences based solely on their alignment. They are implemented in VAMPS [14], TAXONERATOR [15] and CREST [16], or can be performed directly through BLASTN [17] function of the NCBI server (https://blast.ncbi.nlm.nih.gov/Blast.cgi). The performances of CREST and BLASTN are evaluated in detail in this publication. On the other hand, because VAMPS is specifically designed for prokaryotic organisms, while TAXONERATOR uses the same routine as BLASTN, neither one is included in this comparison.

Probabilistic approaches rely on likelihood estimates of OTU placement and include the UTAX algorithm of the USEARCH software package [18] and STATISTICAL ASSIGNMENT PACKAGE (SAP) [19]. For technical reasons, none of these tools are included in this comparison: (i) exact details of the UTAX algorithm have not been published, and thus the results produced by this approach are difficult to evaluate; (ii) a standalone version of SAP could not be successfully installed, while the web server (http://services.birc.au.dk/sap/server) was not stable in use and consistently returned error messages.

The *tree-based* approach evaluates the similarity between query and reference sequences by analysing the position of each individual OTU relative to the reference sequence on the phylogram and the bootstrap support that it receives. This approach includes the following bioinformatic steps: multiple sequence alignment of short query reads with reference sequences is done de novo using any available multiple sequence alignment tool; the dataset is usually trimmed to the barcode size; the phylogram is built using one of the phylogeny inference algorithms, most commonly Neighbour Joining, followed by bootstrapping [20–25].

Phylogeny-based identification of query sequences is performed in three stages. During the preparation stage, a manually curated reference alignment is created using full-length sequences of the gene that includes the barcoding region. A reference phylogeny is estimated based on this alignment. Taxonomic assignment of the query barcodes is then done by using the reference tree as a constraint and testing placement of query reads across all nodes in the reference topology, with the placement likelihood calculated for every combination. The highest scoring placements are retained for evaluation. This approach is implemented in MLTREEMAP [26], PPLACER [27] and Evolutionary Placement Algorithm (EPA) [28]. Of the three, only the EPA is used in this paper, because 'there was no clear difference in accuracy between EPA and PPLACER' (cited from [27]) in comparative tests performed [28]. MLTREEMAP is designed for taxonomy assignment of barcodes into higher-level taxonomic categories (phylum and above) and was not suitable for our purpose.

2. Material and methods

2.1. Sampling sites, sampling, extraction and fixation

Samples used in this study were collected in two ecologically distinct locations along the west coast of Sweden. Coarse shell sand was sampled at 7–8 m depth with a bottom dredge along the northeastern side of the Hållö island near Smögen (N 58° 20.32–20.38′ E 11° 12.73–12.68′). Soft mud was collected using a Warén dredge at 53 m depth in the Gullmarn Fjord near Lysekil (N 58° 15.73′ E 11° 26.10′), in the so-called 'Telekabeln' site. Samples from both sites were extracted using two different techniques each. Material for metabarcoding was preserved in 96% ethanol and stored at −20°C; material for morphology-based identification was preserved in 4% formaldehyde.

The meiofauna from the coarse sand from Hållö was extracted using two variations of the flotation (decanting and sieving) technique. In the first case, fresh water was used to induce osmotic shock in meiofaunal organisms and force them to detach from the substrate. A volume of 200 ml of sediment was placed in a large volume of fresh water, and thoroughly mixed to suspend meiofauna and sediment. The supernatant was sieved through a 1000 µm sieve in order to separate and discard the macrofaunal fraction. The filtered sample was then sieved through a 45 µm sieve to collect the meiofauna, which was preserved either for sequencing or morphological identification. The sieving step was repeated three times. Ten replicates were preserved for molecular studies and two replicates were preserved for morphology-based observations. In the second case, a 7.2% solution of $MgCl_2$ was used to anaesthetize nematodes and other organisms to detach them from the substrate. The meiofauna was decanted through a 125 µm sieve. Similarly, 10 replicates were preserved for molecular studies and two replicates were preserved for morphology-based observations.

The meiofauna from the mud samples was also extracted using two different methods: floatation and siphoning. For the floatation, fresh water was used to induce osmotic shock in meiofaunal organisms. A volume of 2.4 l of sediment was placed in a large volume of fresh water, and thoroughly mixed to suspend the meiofauna and sediment. The supernatant was sieved through a 1000 µm sieve in order to separate and discard the macrofaunal fraction. The filtered sample was then sieved through a 70 µm sieve to collect the meiofauna. The last procedure was repeated three times. The meiofauna was collected, divided into 12 subsamples and preserved: six subsamples were preserved for molecular studies and six subsamples were preserved for morphology-based observations. For siphoning, a total volume of 12 l of sediment was transferred to a plastic container, covered with 20 cm of seawater and left to settle overnight. The meiofauna was then collected through siphoning off the top layer of sediment and passing it through a 125 µm sieve from which samples were taken. One sample was fixed in 96% ethanol, and split into six equal subsamples for molecular studies. The second sample was also split into six subsamples and preserved for morphology-based observations.

2.2. Morphology-based analysis of samples

To estimate nematode diversity, it is usually recommended to count and identify all nematode individuals either in the entire sample or in a subsample of a predetermined volume. The alternative, least time-consuming and most commonly used option is to count a predetermined number (usually 100 or 200) of randomly picked nematodes from the sample. Unfortunately, this latter approach can be imprecise for samples with high species diversity. Moreover, because nematodes are affected by Stokes law, which causes uneven distribution of specimens of different size along the bottom of the counting dish, it is difficult to obtain randomized data with this approach. Therefore, we opted to count and identify all nematodes for all samples (or subsamples). The amount of time required for this task limited the effort to two replicates for each site and extraction method, eight in total. We appreciate that counting nematodes in only two replicates per sample is not enough to quantitatively evaluate the composition of nematode communities; it is nevertheless satisfactory to provide the list of species and genera for each sampling site and extraction method for the purpose of this publication.

All nematode specimens were identified and counted for two replicates each from Hållö floatation with $MgCl_2$, Hållö floatation with fresh water and Telekabeln siphoning. Telekabeln floatation with freshwater was subsampled by taking 1/10 of the entire sample. Specimens from formaldehyde-preserved samples were transferred to pure glycerine using a modified Seinhorst rapid method [29] and mounted on glass slides using the paraffin wax ring method. All nematode specimens were identified to genus and, when possible, to species level and placed in the classification system published in Schmidt-Rhaesa [5] and accepted in WoRMS [30] and NeMys [31] reference databases. Note that this classification

is in many cases different from the nematode classification used in GenBank [32], SILVA [33] and GBIF (www.gbif.org).

2.3. Sequencing procedures

Several different markers are used in barcoding and metabarcoding of biota, including mitochondrial cytochrome c oxidase subunit 1 (COI) [34], ITS rRNA [35], multiple regions of 18S rRNA [1] and 28S rRNA [24,36]. Nematode sequences used in this publication were generated as part of a larger NGS-based meiofauna survey [6], which included sequencing and comparative analysis of both standard animal barcode COI [34] and a marker encompassing a V1–V2 variable region of the 18S rRNA gene originally proposed for barcoding of nematodes [37]. The 18S rRNA sequence was chosen for subsequent analysis for the following reasons: (i) the 18S rRNA (V1–V2) region had a higher sequencing success rate in nematodes with 139 OTUs versus only 22 COI OTUs generated using two different sets of primers [6]; (ii) the reference dataset for marine nematodes includes over 300 high-quality 18S sequences obtainable from GenBank, whereas only about 60 COI barcodes of marine nematodes are available in BOLD; (iii) this particular genetic marker is commonly used in metabarcoding studies of marine meiofauna [2,3,6,38] and plankton [39].

DNA extractions from the samples preserved in 96% ethanol were performed on about 10 g of sediment using the PowerMax® Soil DNA Isolation Kit, (MO BIO Laboratories), according to the manufacturer's instructions. Primers were designed for the 18S rRNA gene including Illumina MiSeq overhang adapter sequences for compatibility with Illumina index and sequencing adapters. The 18S rRNA marker was amplified using PCR primers modified from Fonseca *et al.* [2] yielding an approximately 370 bp fragment that includes the V1–V2 hypervariable domains of 18S rRNA (electronic supplementary material, figure S1). Illumina MiSeq library preparation was done using the dual PCR amplification method [40]. All subsequent sequencing and bioinformatic analysis steps are fully described in Haenel *et al.* [6].

2.4. Preliminary taxonomic assignment using QIIME

Preliminary taxonomic assignment was done using the QIIME [41] script *assign_taxonomy.py* against the SILVA database [33] release 111 in order to identify and separate nematode OTUs from the total of 1472 18S OTUs of meiofauna generated during a previous step [6]. Default settings in QIIME used for preliminary sorting of OTUs grouped query sequences into two groups based on similarity level: to phyla at 80% similarity and to species at 97% similarity. The output for each query sequence included the closest match but did not give the similarity level, making it impossible to evaluate these assignments. Only two OTUs were positively identified using QIIME to species level: *Viscosia viscosa* (TS6.SSU58722) and *Chromadora nudicapitata* (HF2.SSU192072). Six more OTUs were identified to the genus level: *Enoplus* sp. (HE3.SSU110275), *Enoploides* sp. (HE3.SSU124287), *Symplocostoma* sp. (HE5.SSU188855), *Calomicrolaimus* sp. (HF9.SSU20251), *Odontophora* sp. (HF1.SSU779114) and *Sabatieria* sp. (TF6.SSU48167).

The original output from the QIIME analysis included 145 OTUs assigned to the phylum Nematoda. Four of them were incorrectly placed among nematodes due to errors in the reference database derived from SILVA—they group with Arthropoda (HE1.SSU866120, HE6.SSU382930, HF6.SSU331569) and Phoronida (TS6.SSU559982) in all other analyses and were excluded. Two more sequences cluster with nematodes but appear to have long insertions within conserved regions (HE6.SSU358113 and TF5.SSU411806). Both of them were found only in one sample each, further supporting the idea that they are derived from an erroneous amplification product, and were removed from any further analysis. The final list of nematode OTUs includes 139 query sequences.

2.5. Taxonomy assignment of nematode OTUs using alignment-based methods

All 139 nematode OTUs were manually analysed using BLASTN 2.5.0+ [17] against the nucleotide collection of the NCBI database (http://blast.ncbi.nlm.nih.gov/Blast.cgi) on 22 August 2016 with the following settings: *optimize for highly similar sequences* (megablast), *exclude uncultured/environmental sample sequences*, *max target sequences—100*, sorted by max score. Closest matches were evaluated. If the top match sequence was still labelled as 'uncultured', 'unidentified' or 'environmental', the next best match was evaluated. Assignment to the family level was based on the top hit with at least a 90% identity score, with 100% sequence cover, as well as assignment consistency (e.g. top hits assigned to the same family). It is based on a study [42] which defines 99% identity of the 18S rRNA gene equal to species, 96.5% to genera, 90% to families and 84% as equivalent to orders (or 1%, 3.5%, 10% and 16% difference per

position) using single linkage clustering. The chosen threshold was further confirmed by Holovachov [43], who found that a 90% identity score is usually sufficient to assign OTUs (based on V1–V2 region of 18S rRNA) of marine nematodes to families.

The LCAClassifier function of the CREST web server (http://apps.cbu.uib.no/crest) was used to assign taxonomy to 139 OTUs using the built-in silvamod database [16] on 25 August 2016. Three different scores of the LCA relative range were tested separately: 2%, 5% and 10%. The results based on the LCA range of 2% provided the highest number of identified OTUs and were retained for further analysis and comparison.

2.6. Taxonomy assignment of nematode OTUs using tree-based approach

According to published tests [44], the tree-based approach does not allow grouping of sequences into well-supported monophyletic clades equivalent in their taxonomic composition to nematode orders, but most of the marine nematode families are well resolved and supported. The reference sequence dataset was based on the 'filtered' alignment from Holovachov [44] that was updated with newly published sequences of marine nematodes. The final reference dataset is composed of 305 sequences representing the majority of marine nematode families as well as selected freshwater and terrestrial families, some species of which are known to inhabit the marine environment, plus three outgroup taxa (electronic supplementary material, table S1). The same set of sequences was used for the taxonomy placement using a phylogeny-based approach (§2.7).

The reference dataset was trimmed to the barcoding region and aligned with query sequences using the ClustalW [45] algorithm at default settings implemented in MEGA v. 7 [46]. The final alignment was 433 bases long. A phylogenetic tree was built using maximum-likelihood phylogeny inference with RAxML v. HPC2 [47] at default settings (GTR substitution model) with 1000 bootstrap replicates via the CIPRES portal [48]. Two independent analyses were performed: in the first case, all 139 query sequences (cumulative reference dataset) were aligned with the reference dataset and analysed at once; in the second case, 139 query sequences were split into 14 groups of 10 or nine (partitioned query dataset); each group was separately aligned with the complete reference dataset and analysed. This was done to verify if the number and composition of query sequences have any impact on the effectiveness of the tree-based taxonomy assignment approach. OTUs were assigned to the families when they are placed within monophyletic and highly supported clades (bootstrap support of 70% or higher [49,50]), equivalent in their composition to the family-level taxonomic category or below (subfamily, genus), following the same principles that are used when species are classified in supraspecific taxa using the results of phylogenetic analysis [51].

2.7. Taxonomy assignment of nematode OTUs using phylogeny-based approach

Alignments from Holovachov *et al.* [52,53] were combined together and supplemented with other sequences of marine nematodes available in GenBank. To minimize any potential errors and inconsistencies, at the tree-building stage, alignment stage and placement stage, all sequences used for generating reference alignment and the reference tree were selected to be as complete as possible, with the exception of taxa for which no alternative option was available. Secondary structure annotation was manually added to all non-annotated sequences using the JAVA-based editor 4SALE [54], and all sequences were manually aligned to maximize the apparent positional homology of nucleotides. The resulting alignment includes representatives of all families of marine nematodes for which sequence data are available, as well as selected freshwater, terrestrial and animal parasitic taxa (electronic supplementary material, table S1). The reference tree was built using RAxML ver. HPC2 [47] via the CIPRES portal [48] with maximum-likelihood inference of the partitioned dataset. The GTR nucleotide substitution model was used for non-paired sites, whereas the RNA7A [55] substitution model was used for paired sites. Bootstrap maximum-likelihood analysis was performed using the rapid bootstrapping option with 1000 iterations.

Query sequences were aligned to a fixed reference alignment (created in the previous step) using either MOTHUR v. 1.36.1 [56] or PAPARA [57] under default settings. Taxonomy predictions for query sequences were than generated with the EPA [28] implemented in RAxML [47] using the following command: *raxmlHPC-PTHREADS -T 2 -f v -s alignment_file -t reference_tree -m GTRCAT -n output.* Taxonomic assignments to family-level taxonomic categories were based either on high likelihood (above the 95% threshold) of a single placement, or on high cumulative likelihood (above the 95% threshold) of multiple placements, all of which are within a single strongly supported monophyletic clade equal to a family (see §4.4 for explanation). The 95% similarity threshold is the default used by the EPA.

2.8. Image processing

Trees were visualized using FigTree [58] and iTOL [59]. All clades with bootstrap support lower than 70% were collapsed in the final illustrations. Secondary structure of the barcoding region of 18S rRNA (electronic supplementary material, figure S1) was visualized using VARNA [60].

3. Results

3.1. Morphology-based analysis of samples

The nematode fauna in the coarse sand from the Hållö site included 107 different nematode species belonging to 86 genera and 33 families (electronic supplementary material, table S2). Of these, floatation using $MgCl_2$ recovered 88 species from 73 genera and 26 families, while floatation using H_2O recovered 101 species from 83 genera and 33 families. The differences in nematode fauna extracted using two variations of the same method are limited to rare species of different size classes (from less than 0.5 mm to over 2 mm). Relative abundance of these rare species does not exceed 0.14% (0.01–0.14%, with an average of 0.03%). The list of nematodes from the Hållö site includes four species new to the fauna of Sweden (*Bolbonema brevicolle*, *Bradylaimus pellita*, *Desmodora granulata* and *Odontophora villoti*) and five species new to science (from the genera *Adelphos*, *Paramesonchium*, *Leptolaimus* and *Diplopeltoides*).

Mud sediments from the Telekabeln site were inhabited by 113 different nematode species, belonging to 77 genera and 33 families (electronic supplementary material, table S3). Of these, siphoning recovered 81 species from 62 genera and 29 families, while floatation using H_2O recovered 102 species from 70 genera and 32 families. The differences in nematode fauna extracted using two different methods include both rare and uncommon species of various size classes (from less than 0.5 mm to over 2 mm). The relative abundance of these rare species does not exceed 2.02% (0.01–2.02%, with an average of 0.29%). The list of nematodes from the Telekabeln samples includes seven species new to the fauna of Sweden (*Campylaimus rimatus*, *C. amphidialis*, *C. tkatchevi*, *C. orientalis*, *Diplopeltoides asetosus*, *D. linkei* and *D. nudus*) and one species new to science (from the genus *Diplopeltoides*).

3.2. Taxonomy placement of OTUs using alignment-based approaches

3.2.1. BLASTN

Out of 139 queried OTUs, 52 could be assigned to family-level categories based on the following criteria: 90% or more identity score and 100% sequence cover, as well as assignment consistency (electronic supplementary material, table S4). In one case, BLASTN search produced conflicting results—two top hits with the same identity score and sequence cover that belonged to different families, but still falling within the threshold limit. This is the barcode TF1.SSU676746 that showed 90% identity and 100% sequence cover to *Haliplectus* sp. (family Haliplectidae) and *Prodesmodora* sp. (family Desmodoridae). It was considered unassigned. Similar examples were seen in BLASTN results of other OTUs that did not reach the threshold. These examples show that considering only one top hit when assigning taxonomy to query OTUs using alignment-based approaches may sometimes lead to questionable or dubious identification.

3.2.2. CREST

Only 26 out of 139 queried OTUs were assigned to families using LCAClassifier of CREST under default parameters (electronic supplementary material, table S5) and following built-in classification. In two cases, OTUs were placed outside Nematoda: HE3.SSU118424 was placed within Copepoda (phylum Arthropoda) and TS1.SSU284163 was placed in Scolecida (phylum Annelida). The first OTU was positively assigned to the family Oxystominidae (phylum Nematoda) using tree-based and phylogeny-based approaches (see §3.3 and 3.4); the second OTU was unassigned in all other analyses.

3.3. Taxonomy placement of OTUs using tree-based approaches

3.3.1. Cumulative query dataset

Tree-based taxonomy assignment of the cumulative query dataset produced 54 well-supported placements (figure 1; electronic supplementary material, table S6) that fulfilled the following criteria:

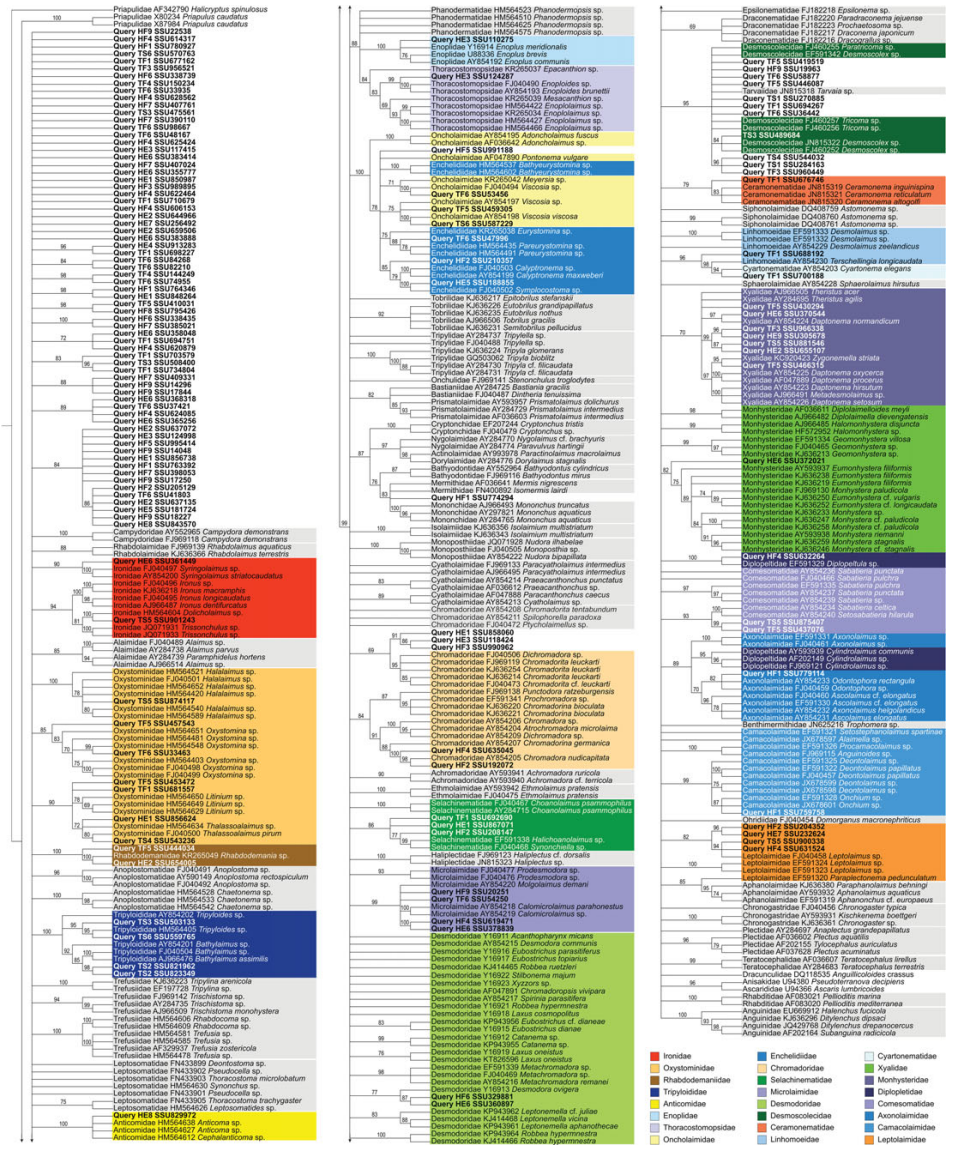


Figure 1. Phylogram based on tree-based taxonomy assignment approach using a complete query dataset. Families that include positively assigned OTUs are colour-coded; remaining reference taxa are shaded in grey.

OTU must cluster within the monophyletic clade that has high bootstrap support (greater than or equal to 70%) and is at or below family level. The remaining 85 OTUs could not be placed in clades satisfying these criteria, and are thus treated as unidentified.

3.3.2. Partitioned query dataset

The results of taxonomic assignment using a tree-based approach of the partitioned query dataset produced somewhat different results compared to the cumulative query dataset—67 OTUs were placed in monophyletic clades equivalent to family-level categories with sufficient support (electronic supplementary material, table S6). Of these, taxonomic placement of only 47 OTUs matched the identification produced using the cumulative query dataset, and identifications of 20 OTUs were new. Seven OTUs were not assigned using a partitioned query dataset but were positively identified using a cumulative query dataset.

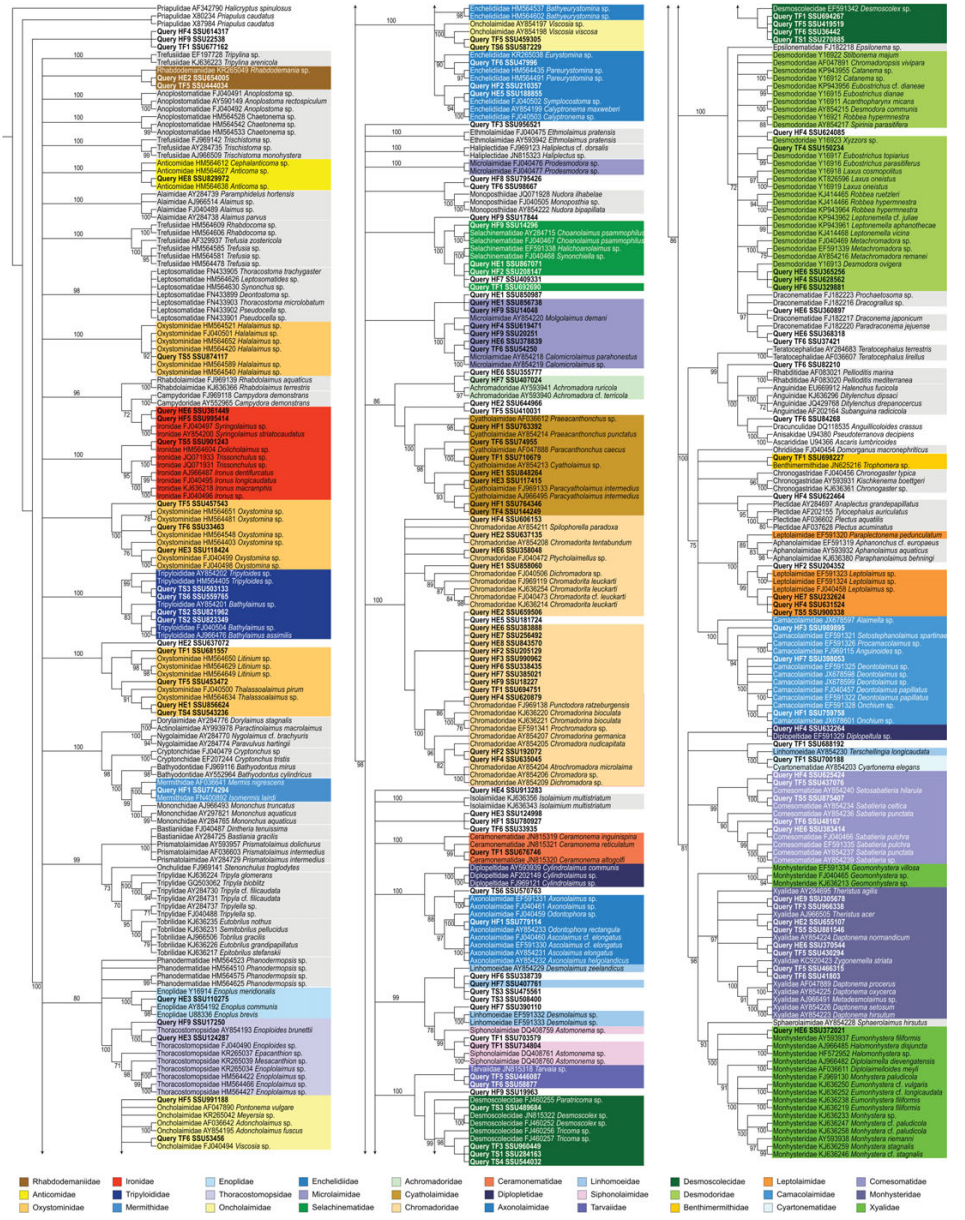


Figure 2. Phylogram based on phylogeny-based taxonomy assignment approach. Families that include positively assigned OTUs are colour-coded; remaining reference taxa are shaded in grey.

3.4. Taxonomy placement of OTUs using phylogeny-based approaches

3.4.1. EPA/MOTHUR

Phylogeny-based taxonomy assignment using MOTHUR-based alignment and the EPA produced 105 well-supported placements with single or accumulated likelihood of 0.95 or more (figure 2; electronic supplementary material, table S7). There are ten additional cases when the positive identity cannot be attained because OTUs are placed either within a paraphyletic assemblage (family Desmodoridae or Linhomoeidae) or closely related monophyletic clade (Draconematidae or Siphonolaimidae, respectively).

3.4.2. EPA/PaPaRa

The results produced using PaPaRa-based alignment and the EPA are exactly the same as those obtained using mothur-based alignment and described in §3.4.1 (electronic supplementary material, table S7), even though visual comparison of alignments produced by mothur and by PaPaRa revealed some differences.

3.5. Comparison of different taxonomy assignment approaches

Among the three different taxonomy assignment approaches tested (each with two variations), the EPA (both variations) placed the largest number of query OTUs into family-level taxonomic categories (105 out of 139), while CREST implementation of the alignment-based assignment approach was the least efficient (26 out of 139). Despite such a broad success rate, the family identifications were in most cases congruent among different approaches—most of the identified OTUs were assigned to the same families (electronic supplementary material, table S8), with the following exceptions:

(i) HF1.SSU759758 was placed in the family Camacolaimidae using tree-based and phylogeny-based approaches, in the family Leptolaimidae using CREST and unassigned using BLASTN;
(ii) HF5.SSU995414 was placed in the family Rhabdolaimidae using BLASTN, in the family Ironidae using CREST and both variations of the EPA, and unassigned using the tree-based approach;
(iii) TF1.SSU698227 was placed in the family Teratocephalidae using BLASTN and in the family Benthimermithidae using both variations of the EPA, and unassigned in other cases;
(iv) TF1.SSU700188 was placed in the family Linhomoeidae using BLASTN, in the family Cyartonematidae using tree-based and phylogeny-based approaches, and unassigned using CREST;
(v) TF6.SSU47996 was placed in the family Oncholaimidae using BLASTN and in the family Enchelidiidae in all other cases.

3.6. Comparison between barcode-based and morphology-based identification

The EPA (phylogeny-based approach) provided the largest number of positively identified OTUs and will be compared with the faunistic lists created by identifying nematode specimens using morphological characters. As species-level identification cannot be achieved for most of the OTUs, the results of barcode-based and morphology-based identifications can only be compared as the number of identified OTUs/morphospecies per family (figure 3; electronic supplementary material, table S9). Among families with available reference sequences, barcode-based identification failed to identify the families Phanodermatidae, Leptosomatidae, Trefusiidae, Epsilonematidae, Draconematidae, Monoposthiidae and Sphaerolaimidae. One of the likely explanations is that nematodes from these families failed to amplify or that barcode sequences produced during sequencing failed quality control.

On the other hand, barcode-based identification also uncovered several taxa that were overlooked during morphology-based identification, such as the families Achromadoridae, Mermithidae and Benthimermithidae—the last two are internal parasites of invertebrates during part of their life cycle and were most probably overlooked, because examination of the meiofauna for internal parasites was not attempted. In all other cases, the efficiency of either barcode-based or morphology-based identification varied considerably, even within the same taxon across different samples (figure 3). Nevertheless, the Pearson correlation coefficient revealed moderate positive correlation ($\rho = 0.7296967138$) between the number of assigned OTUs and identified morphospecies in each family/extraction/sample (electronic supplementary material, figure S2).

4. Discussion

4.1. General notes

Three different taxonomy assignment approaches (with two modifications each) tested in this project provide some variation in the number of positively identified OTUs; however, the assigned identities of

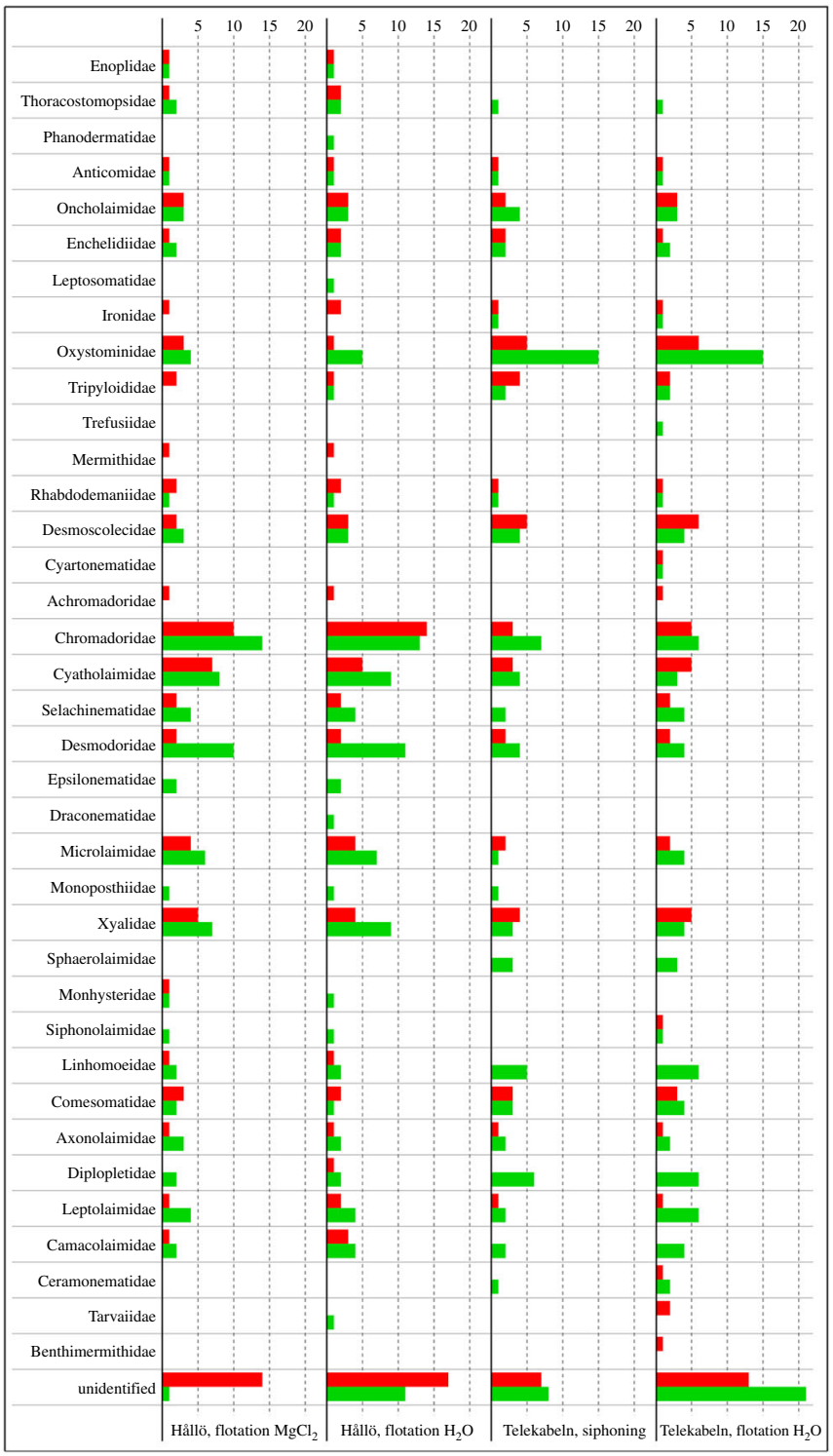


Figure 3. Comparison of the total number of taxa identified using phylogeny-based taxonomy assignment approach (OTUs, red) and morphology-based identification (morphospecies, green) for each nematode family in each sample (sampling site/extraction method) based on table S9 in the electronic supplementary material (excluding families without reference sequence data).

those OTUs that were identified were consistent with very few exceptions (§3.5). These discrepancies can possibly be caused by several different factors. Placement of one of the OTUs (HF1.SSU759758) either in the family Camacolaimidae (tree-based and phylogeny-based approaches) or in the family Leptolaimidae (CREST) is probably a result of outdated classification of the phylum Nematoda used in the SILVA-derived reference database implemented in CREST, compared to the nematode classification used in WoRMS and in this publication (§4.6). Conflicting results of the assignment of TF1.SSU698227 either in the family Teratocephalidae (BLASTN) or in the family Benthimermithidae (EPA) can be due to poor representation of the reference dataset in this part of the nematode tree. The remaining conflicting placements of HF5.SSU995414 (Rhabdolaimidae versus Ironidae), TF1.SSU700188 (Linhomoeidae versus Cyartonematidae) and TF6.SSU47996 (Oncholaimidae versus Enchelidiidae) are possibly caused by the fact that the overall sequence similarity used by BLASTN does not necessarily reflect common phylogenetic history, which is the basis of the tree-based and phylogeny-based assignment approaches. Differences in the individual success rates of each taxonomy assignment approach will be discussed in §4.2–4.4.

4.2. Alignment-based approach

Alignment-based approaches tested in this publication include manual analysis using BLASTN 2.5.0+ [17] against the nucleotide collection of the NCBI database and the LCAClassifier function of the CREST against the built-in silvamod database [16]. Both tested approaches have their own advantages and disadvantages. NCBI implementation of BLASTN allows visual examination of multiple top hits in the output and individual evaluation of these top hits, manual application of the variable similarity threshold if it has been predetermined empirically and, if necessary, correction of classification. Taxonomy assignment using CREST is less flexible and has the following limitations: (i) similarity thresholds used in CREST are based on the prokaryotic 16S rRNA analysis and do not account for the differences in the variability of rRNA within and between different taxa [43]; (ii) classification of the phylum Nematoda that is used in the CREST database is different from the most recent and widely accepted classification scheme published in WoRMS; and (iii) results of the taxonomy assignment in the output files cannot be verified and, if necessary, updated.

Strictly speaking, alignment-based assignment approaches should not be used to place OTUs to supraspecific taxa without critical evaluation of the results. First of all, similarity scores used in BLASTN search results do not reflect phylogenetic affinities of analysed taxa, and do not account for the fact that the level of variability of the 5′ barcoding region of 18S rRNA (electronic supplementary material, figure S1) is different in various nematode taxa [43]. Too narrow similarity thresholds can exclude potentially identifiable sequences, while too broad thresholds can lead to misidentifications. Dell'Anno *et al*. [4] is an example where broad similarity threshold resulted in incorrect assignment of several nematode OTUs from deep-sea samples to nematode species known to inhabit freshwater and soil and never found in the marine environment (e. g. *Anaplectus porosus*, *Anaplectus* sp., *Pakira orae* and *Tylolaimophorus* sp.).

4.3. Tree-based approach

Phylogenetic hypotheses used to infer relationships of taxa are usually thoroughly described and rigorously evaluated, and undergo comparison and testing using different alignment and tree-building algorithms. Phylogenetic trees used to identify unknown barcodes are less so [20,21]. Barcodes are by definition relatively short in length, hypervariable sites flanked by conserved regions. Hypervariable domains V1 and V2, which are part of the barcoding region of the 18S rRNA used in this publication, are the culprit that causes poor alignment and hence has negative effect on the quality of the resulting phylogeny. Different alignment and phylogeny-inference algorithms may provide competing phylogenetic hypotheses [44] and, as a result, different placements of OTUs in the phylogram. Taxon composition and sequence quality (exclusion of incorrectly identified species, low quality and short sequences) of the reference dataset is also crucial [44], as it determines which taxa can be identified and which taxa cannot. Even the number and composition of OTUs have strong effect on the final phylogenetic tree and, as a result, on the outcome of the taxonomy assignment, as shown in §3.3. The latter is caused by the need to align de novo the combined datasets that include reference and query sequences—the presence of unidentified sequencing errors among query OTUs can have a negative effect on the alignment and phylogeny inference, even if all reference sequences are of high quality. This effect is global, i.e. by affecting the entire alignment and tree topology and bootstrap, erroneous sequences can potentially cause other OTUs to be misidentified or unidentified. In conclusion, successful use of

tree-based approaches to assign taxonomy to OTUs is highly dependent not only on the quality and completeness of the reference dataset and alignment and phylogeny inference algorithms, but also on the quality and diversity of query sequences.

4.4. Phylogeny-based approach

Phylogeny-based approaches allow the estimation of the most likely position of each OTU within the constrained phylogenetic tree, estimation of the rank of its taxonomic placement in supraspecific categories if these are well resolved and supported in the reference phylogeny, and can even work with paraphyletic taxa. Moreover, because the reference alignment and reference phylogeny are constrained during phylogeny-based taxonomy assignment procedures, the quality of query sequences has no impact on the result, i.e. the presence of erroneous sequences among query OTUs (chimaeras) has no effect on the identification of other query OTUs. The outcome of the analysis solely depends on the quality of the reference alignment and reference phylogeny. Even minor differences in the alignment of OTUs against the reference alignment noted above (§3.4.2) had no effect on the results. An additional advantage of the phylogeny-based taxonomy assignment approach implemented in the EPA is the possibility to use cumulative likelihood scores when assigning taxa to clades equivalent to supraspecific taxonomic categories (electronic supplementary material, figure S3).

4.5. Metabarcoding versus morphology-based identification

Morphology-based identification procedures are strongly biased by the expertise and experience of the researcher performing the identification, as well as the state of the knowledge on the diversity of particular groups of nematodes. Metabarcoding, on the other hand, should be able to better estimate the diversity of poorly known groups of nematodes, or groups for which taxonomic expertise is not available at the moment, as well as unidentifiable specimens (eggs, juveniles, damaged specimens, etc.). Moreover, metabarcoding can reveal taxa that are physically hidden and cannot be observed by the researcher during sorting and identification, such as internal parasites—similarly to the results obtained by Lindeque *et al*. [39], barcode-based identification revealed the presence of endoparasitic nematodes from the families Mermithidae and Benthimermithidae in our samples. They had been overlooked during morphology-based identification, probably being juveniles within bodies of other invertebrates.

The number of OTUs identified by metabarcoding is strongly influenced by the clustering procedures of the raw sequence data and, depending on the threshold used, will give different results. Assuming that the OTUs produced through metabarcoding are equivalent to currently recognized morphospecies, the only reason it would not be able to correctly estimate the number of species in the sample is if there are issues with amplification of the barcoding gene. The genus *Halalaimus* is a good example of a problematic taxon in this case—only one *Halalaimus* OTU (TS5.SSU874117) was recovered with metabarcoding, and only from the Telekabeln site. Morphology-based identification recovered at least two different *Halalaimus* species in the Hållö site and more than eight species in the Telekabeln site, some of which were relatively common. GenBank hosts a number of *Halalaimus* sequences, confirming that the genus is sufficiently diverse genetically, and that our single *Halalaimus* OTU is unlikely to encompass multiple morphospecies, but is rather a result of amplification problems.

4.6. Reference databases

Taxonomy assignment procedures described in the literature [16,41] often rely on various releases of the SILVA database [33], which in turn is based on the sequence data published in GenBank or EMBL. These databases can be 'built-in' (CREST), and completely inaccessible for the user, or 'pre-made' and hard to modify (QIIME). The presence of erroneously identified sequences of nematodes and other organisms in GenBank and SILVA databases has been mentioned multiple times [43,44,61,62]. If the reference database is not checked for errors prior to the analysis, the results produced by any taxonomy assignment algorithm should be evaluated using available data on geographical or ecological distribution of species, in order to avoid mistakes.

As mentioned earlier, the SILVA database in itself does not always follow the most recent accepted classification for certain groups of organisms. As a result, placing some of the OTUs into nematode families based on the SILVA classification turned out to be incorrect. For example, genera *Paracyatholaimus* and *Preacanthonchus* were placed in the family Chromadoridae using QIIME, while

they do belong to the family Cyatholaimidae. Similar examples are: *Enoploides* placed in Enoplidae instead of Thoracostomopsidae, *Calyptronema* in Oncholaimidae instead of Enchelidiidae, *Achromadora* in Chromadoridae instead of Achromadoridae, *Camacolaimus* in Leptolaimidae instead of Camacolaimidae, and some others. Output from CREST [162] only gives the name of the supraspecific taxon for those cases where a query OTU cannot be identified to species level. This prevents proper evaluation of the assignment results and correction of assignments derived from an erroneous reference sequence or incorrect classification. We do not expect any database to be able to quickly reflect changes in nematode classification, but we expect end users of these databases to be aware of the need to verify and, if necessary, to update the output of any taxonomy assignment procedure that they may use.

Another disadvantage of taxonomy assignment software that uses built-in databases and offers only top-pick assignments in the output files (QIIME, CREST) is that a substantial number of OTUs are matched with environmental samples, labelled in such databases with the words 'environmental' (e.g. 'environmental sample'), 'uncultured' (e.g. 'uncultured eukaryote') and 'unidentified' ('unidentified nematode'). They themselves are OTUs generated during previous metabarcoding projects and identified not by looking at actual morphological vouchers but by using one of the multiple taxonomy assignment methods. Moreover, by giving only one top 'hit' assignment, such software eliminates the possibility to verify if the 'second best' hit is based on sequence data from the physically observed and identified morphological voucher, and its similarity score, preventing the researcher from making educated decisions on the taxonomic identity of an OTU.

5. Conclusion and future prospects

The identification of OTUs is obviously a key step in metabarcoding and it is essential that the most effective method is used (as opposed to the fastest or simplest). Ideally, the barcode sequences should be assigned taxonomic names that provide a link to all biological knowledge that may exist in relation to the organism. Misidentification will compromise the results, for example, in studies of biogeography, community structure, habitat state or the presence of certain important species (invasive, rare, indicators, etc.).

Identification of OTUs should be at the appropriate taxonomic level, which is determined by the available reference sequences and the purpose of the study. In the case of marine nematodes, we were able to assign our barcode sequences to family-level taxa to a high degree despite the very incomplete reference database. The relevance of family-level metabarcoding data in ecological studies remains poorly tested and requires extensive comparison with data obtained using classical approaches.

The full potential of metabarcoding is realized when sequences are identified to species level. This conveys the most information and permits more robust inferences. A prerequisite for this is taxonomic groundwork in the form of complete curated reference databases with sequences of reliably identified specimens.

We found the phylogeny-based taxonomy assignment approach to be the most efficient and the least error-prone. The alignment-based approach is less reliable because the similarity thresholds it depends on do not account for inter- and intra-taxon variations in barcode sequence, while tree-based approaches can be affected by the quality of the input OTU data. If phylogeny-based taxonomy assignment methods become widely used in nematode metabarcoding, it is imperative to create and maintain high-quality reference alignments and reference phylogenetic trees to be used by researchers worldwide.

Ethics. There are no particular ethical aspects specific to this publication. It did not involve: (i) experiments on animals, (ii) collection of protected species, (iii) research on human subjects or (iv) collection of personal data.

Data accessibility. The data supporting this article are available in the electronic supplementary material.

Authors' contributions. O.H. conceived and designed the study, performed morphology-based identification and taxonomy assignment analyses and wrote the manuscript with input from Q.H., S.J.B. and U.J. U.J. and O.H. performed fieldwork. Q.H. and S.J.B. performed molecular analyses. Q.H., O.H., U.J. and S.J.B. contributed reagents, materials and analysis tools. All authors gave their final approval for publication.

Competing Interests. We declare we have no competing interests.

Funding. This work was in part supported by the project 'Systematics of Swedish free-living nematodes of the orders Desmodorida and Araeolaimida' (Swedish Taxonomy Initiative, ArtDatabanken, Sweden) awarded to O.H. and by the Swedish Research Council project (2012-3446) 'Biodiversity genomics: Species identification pipelines for analysing marine invertebrate larval stages, community structure, and trophic interactions' awarded to S.J.B.

Acknowledgements. We thank the Genomics Core facility platform at the Sahlgrenska Academy, University of Gothenburg. Sampling was conducted using the vessel (Oscar von Sydow) and facilities of the Sven Lovén Centre for Marine Infrastructure in Kristineberg.

References

1. Leray M, Knowlton N. 2016 Censusing marine eukaryotic diversity in the twenty-first century. *Phil. Trans. R. Soc. B* **371**, 20150331. (doi:10.1098/rstb.2015.0331)
2. Fonseca VG *et al.* 2010 Second-generation environmental sequencing unmasks marine metazoan biodiversity. *Nat. Comm.* **1**, 98. (doi:10.1038/ncomms1095)
3. Fonseca VG *et al.* 2014 Metagenetic analysis of patterns of distribution and diversity of marine meiobenthic eukaryotes. *Glob. Ecol. Biogeogr.* **23**, 1293–1302. (doi:10.1111/geb.12223)
4. Dell'Anno A, Carugati L, Corinaldesi C, Riccioni G, Danovaro R. 2015 Unveiling the biodiversity of deep-sea nematodes through metabarcoding: are we ready to bypass the classical taxonomy? *PLoS ONE* **10**, e0144928. (doi:10.1371/journal.pone.0144928)
5. Schmidt-Rhaesa A. 2014 *Handbook of zoology. Gastrotricha, cycloneuralia and gnathifera. Volume 2. Nematoda*. Berlin, Germany: De Gruyter.
6. Haenel Q, Holovachov O, Jondelius U, Sundberg P, Bourlat SJ. 2017 NGS-based biodiversity and community structure analysis of meiofaunal eukaryotes in shell sand from Hällö island, Smögen, and soft mud from Gullmarn Fjord, Sweden. *Biodivers. Data J.* **5**, e12731. (doi:10.3897/BDJ.5.e12731)
7. Jensen P. 1987 Feeding ecology of free-living aquatic nematodes. *Mar. Ecol. Prog. Ser.* **35**, 187–196. (doi:10.3354/meps035187)
8. Yeates GW, Bongers T, de Goede RGM, Freckman DW, Georgieva SS. 1993 Feeding habits in soil nematode families and genera—an outline for soil ecologists. *J. Nematol.* **25**, 315–331.
9. Bongers T, Bongers M. 1998 Functional diversity of nematodes. *App. Soil Ecol.* **10**, 239–251. (doi:10.1016/S0929-1393(98)00123-1)
10. Bongers T. 1999 The Maturity Index, the evolution of nematode life history traits, adaptive radiation and cp-scaling. *Plant Soil* **212**, 13–22. (doi:10.1023/A:1004571900425)
11. Ahmed M, Sapp M, Prior T, Karssen G, Back MA. 2016 Technological advancements and their importance for nematode identification. *Soil* **2**, 257–270. (doi:10.5194/soil-2-257-2016)
12. Kerfahi D, Tripathi BM, Porazinska DL, Park J, Go R, Adams JM. 2016 Do tropical rain forest soils have greater nematode diversity than high Arctic tundra? A metagenetic comparison of Malaysia and Svalbard. *Glob. Ecol. Biogeogr.* **25**, 716–728. (doi:10.111/geb.12448)
13. Kerfahi D, Park J, Tripathi BM, Singh D, Porazinska DL, Moroenyane I, Adams JM. 2017 Molecular methods reveal controls on nematode community and unexpectedly high nematode diversity, in Svalbard high Arctic tundra. *Polar Biol.* **40**, 765–776. (doi:10.1007/s00300-016-1999-6)
14. Sogin ML, Morrison HG, Huber JA, Welch DM, Huse SM, Neal PR, Arrieta JM, Herndl GJ. 2006 Microbial diversity in the deep sea and the underexplored 'rare biosphere'. *Proc. Natl Acad. Sci. USA* **103**, 12 115–12 120. (doi:10.1073/pnas.0605127103)
15. Jones M, Ghoorah A, Blaxter M, Poon AFY. 2011 jMOTU and Taxonerator: turning DNA barcode sequences into annotated operational taxonomic units. *PLoS ONE* **6**, e19259. (doi:10.1371/journal.pone.0019259)
16. Lanzén A, Jørgensen S, Huson D, Gorfer M, Grindhaug SH, Jonassen I, Øvreås L, Urich T. 2012 CREST—classification resources for environmental sequence tags. *PLoS ONE* **7**, e49334. (doi:10.1371/journal.pone.0049334)
17. Madden T. 2002 The BLAST sequence analysis tool, Ch. 16. In *The NCBI handbook*. Bethesda, MD: National Center for Biotechnology Information.
18. Edgar RC. 2010 Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460–2461. (doi:10.1093/bioinformatics/btq461)
19. Munch K, Boomsma W, Huelsenbeck J, Willerslev E, Nielsen R. 2008 Statistical assignment of DNA sequences using Bayesian phylogenetics. *Syst. Biol.* **57**, 750–757. (doi:10.1080/10635150802422316)
20. Morise H, Miyazaki E, Yoshimitsu S, Eki T, Balcazar JL. 2012 Profiling nematode communities in unmanaged flowerbed and agricultural field soils in Japan by DNA barcode sequencing. *PLoS ONE* **7**, e51785. (doi:10.1371/journal.pone.0051785)
21. Sapkota R, Nicolaisen M. 2015 High-throughput sequencing of nematode communities from total soil DNA extractions. *BMC Ecol.* **15**, 3 (doi:10.1186/s12898-014-0034-4)
22. Bhadury P, Austen M, Bilton D, Lambshead P, Rogers A, Smerdon G. 2006 Development and evaluation of a DNA-barcoding approach for the rapid identification of nematodes. *Mar. Ecol. Prog. Ser.* **320**, 1–9. (doi:10.3354/meps320001)
23. Bhadury P, Austen M. 2010 Barcoding marine nematodes: an improved set of nematode 18S rRNA primers to overcome eukaryotic co-interference. *Hydrobiologia* **641**, 245–251. (doi:10.1007/s10750-009-0088-z)
24. De Ley P *et al.* 2005 An integrated approach to fast and informative morphological vouchering of nematodes for applications in molecular barcoding. *Phil. Trans. R. Soc. B* **360**, 1945–1958. (doi:10.1098/rstb.2005.1726)
25. Derycke S, Vanaverbeke J, Rigaux A, Backeljau T, Moens T, Roopnarine P. 2010 Exploring the use of cytochrome oxidase c subunit 1 (COI) for DNA barcoding of free-living marine nematodes. *PLoS ONE* **5**, e13716. (doi:10.1371/journal.pone.0013716)
26. Stark M, Berger SA, Stamatakis A, von Mering C. 2010 MLTreeMap—accurate Maximum Likelihood placement of environmental DNA sequences into taxonomic and functional reference phylogenies. *BMC Genomics* **11**, 461. (doi:10.1186/1471-2164-11-461)
27. Matsen FA, Kodner RB, Armbrust EV. 2010 pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a xed reference tree. *BMC Bioinform.* **11**, 538. (doi:10.1186/1471-2105-11-538)
28. Berger SA, Krompass D, Stamatakis A. 2011 Performance, accuracy, and web server for evolutionary placement of short sequence reads under maximum likelihood. *Syst. Biol.* **60**, 291–302. (doi:10.1093/sysbio/syr010)
29. Grisse AT D. 1969 Redescription ou modifications de quelques techniques utilisées dans l'etude des nematodes phytoparasitaires. *Meded. Fac. Landbouwwet. Gent.* **34**, 351–369.
30. WoRMS Editorial Board. 2016 World register of marine species. See http://www.marinespecies.org
31. Guilini K *et al.* 2016 NeMys: world database of free-living marine nematodes. See http://nemys.ugent.be
32. Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. 2013 GenBank. *Nucleic Acids Res.* **41**, D36–D42. (doi:10.1093/nar/gks1195)
33. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glockner FO. 2013 The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* **41**, D590–D596. (doi:10.1093/nar/gks1219)
34. Hebert PDN, Ratnasingham S, de Waard JR. 2003 Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. *Proc. R. Soc. Lond. B* **270**, S96–S99. (doi:10.1098/rsbl.2003.0025)
35. Seifert KA. 2009 Progress towards DNA barcoding of fungi. *Mol. Ecol. Resour.* **9**, 83–89. (doi:10.1111/j.1755-0998.2009.02635.x)
36. Schenk J, Hohberg K, Helder H, Ristau K, Traunspurger W. In press. The D3-D5 region of large subunit ribosomal DNA provides good resolution of German limnic and terrestrial nematode communities. *Nematology*.
37. Floyd R, Abebe E, Papert A, Blaxter M. 2002 Molecular barcodes for soil nematode identification. *Mol. Ecol.* **11**, 839–850. (doi:10.1046/j.1365-294X.2002.01485.x)
38. Creer S *et al.* 2010 Ultrasequencing of the meiofaunal biosphere: practice, pitfalls and promises. *Mol. Ecol.* **19**, 4–20. (doi:10.1111/j.1365-294X.2009.04473.x)
39. Lindeque PK, Parry HE, Harmer RA, Somerfield PJ, Atkinson A, Ianora A. 2013 Next generation sequencing reveals the hidden diversity of zooplankton assemblages. *PLoS ONE* **8**, e81327. (doi:10.1371/journal.pone.0081327)
40. Bourlat SJ, Haenel Q, Finnman J, Leray M. 2016 Preparation of amplicon libraries for metabarcoding of marine eukaryotes using Illumina MiSeq: the Dual-PCR method. In *Marine genomics - methods and protocols* (ed. SJ Bourlat), pp. 197–208. Berlin, Germany: Springer.
41. Caporaso JG *et al.* 2010 QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods* **7**, 335–336. (doi:10.1038/nmeth.f.303)
42. Cole JR, Konstantinidis K, Farris RJ, Tiedje JM. 2010 Microbial diversity and phylogeny: extending from rRNAs to genomes. In *Environmental molecular microbiology* (eds WY Liu, JK Jansson), pp. 1–20. Norfolk, UK: Caister Academic Press.
43. Holovachov O. 2016 Metabarcoding of marine nematodes—evaluation of similarity scores used in alignment-based taxonomy assignment approach. *Biodivers. Data J.* **4**, e10647. (doi:10.3897/BDJ.4.e10647)
44. Holovachov O. 2016 Metabarcoding of marine nematodes—evaluation of reference datasets used in tree-based taxonomy assignment approach. *Biodivers. Data J.* **4**, e10021. (doi:10.3897/BDJ.4.e10021)

45. Larkin MA *et al.* 2007 Clustal W and Clustal X version 2.0. *Bioinformatics* **23**, 2947–2948. (doi:10.1093/bioinformatics/btm404)
46. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. 2013 MEGA6: molecular evolutionary genetics analysis Version 6.0. *Mol. Biol. Evol.* **30**, 2725–2729. (doi:10.1093/molbev/mst197)
47. Stamatakis A. 2014 RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313. (doi:10.1093/ bioinformatics/btu033)
48. Miller MA, Pfeier W, Schwartz T. 2010 Creating the CIPRES science gateway for inference of large phylogenetic trees. In *Proc. of the Gateway Computing Environments Workshop (GCE), 14 Nov 2010, New Orleans, LA, USA*, pp. 1–8.
49. Hills DM, Bull JJ. 1993 An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Syst. Biol.* **42**, 182–192. (doi:10.1093/sysbio/42.2.182)
50. Soltis PS, Soltis DE. 2003 Applying the bootstrap in phylogeny reconstruction. *Stat. Sci.* **18**, 256–267. (doi:10.1214/ss/1063994980)
51. Wiley EO, Lieberman BS. 2011 *Phylogenetics: theory and practice of phylogenetic systematics*, 2nd edn. Hoboken, NJ: Wiley-Blackwell.
52. Holovachov O, Rodrigues CF, Zbinden M, Duperron S. 2013 *Trophomera conchicola* sp. n. (Nematoda: Benthimermithidae) from chemosymbiotic bivalves *Idas modiolaeiformis* and *Lucionoma kazani* (Mollusca: Mytilidae and Lucinidae) in Eastern Mediterranean. *Russ. J. Nematol.* **21**, 1–12.
53. Holovachov O, Boström S, Tandingan De Ley I, Robinson C, Mundo-Ocampo M, Nadler SA. 2013 Morphology, molecular characterisation and systematic position of the genus *Cynura* Cobb, 1920 (Nematoda: Plectida). *Nematology* **15**, 611–627. (doi:10.1163/15685411-00002706)
54. Seibel PN, Müller T, Dandekar T, Schultz J, Wolf M. 2006 4SALE—A tool for synchronous RNA sequence and secondary structure alignment and editing. *BMC Bioinform.* **7**, 498. (doi:10.1186/1471-2105-7-498)
55. Higgs PG. 2000 RNA secondary structure: physical and computational aspects. *Q. Rev. Biophys.* **33**, 199–253. (doi:10.1017/S0033583500003620)
56. Schloss PD *et al.* 2009 Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.* **75**, 7537–7541. (doi:10.1128/AEM.01541-09)
57. Berger SA, Stamatakis A. 2011 Aligning short reads to reference alignments and trees. *Bioinformatics* **27**, 2068–2075. (doi:10.1093/bioinformatics/btr320)
58. Rambaut A. 2015 FigTree. See http://tree.bio.ed.ac.uk/software/gtree/
59. Letunic I, Bork P. 2016 Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res.* **44**, W242–W245. (doi:10.1093/nar/gkw290)
60. Darty K, Denise A, Ponty Y. 2009 VARNA: interactive drawing and editing of the RNA secondary structure. *Bioinformatics* **25**, 1974–1975. (doi:10.1093/bioinformatics/btp250)
61. Buhay JE. 2009 'COI-Like' sequences are becoming problematic in molecular systematic and DNA barcoding studies. *J. Crustacean Biol.* **29**, 96–110 (doi:10.1651/08-3020.1)
62. Schnell IB, Sollmann R, Calvignac-Spencer S, Siddall ME, Yu DW, Wilting A, Gilbert MTP. 2015 iDNA from terrestrial haematophagous leeches as a wildlife surveying and monitoring tool—prospects, pitfalls and avenues to be developed. *Front. Zool.* **12**, 302. (doi:10.1186/s12983-015-0115-z)

Chapter 7

DNA metabarcoding reveals diverse diet of the three-spined stickleback in a coastal ecosystem

Jakubavičiūtė et al. 2017, PLoS ONE

PLOS ONE

RESEARCH ARTICLE

DNA metabarcoding reveals diverse diet of the three-spined stickleback in a coastal ecosystem

Eglė Jakubavičiūtė[1]*, Ulf Bergström[2], Johan S. Eklöf[3], Quiterie Haenel[4], Sarah J. Bourlat[5]

1 Nature Research Centre, Vilnius, Lithuania, **2** Department of Aquatic Resources, Institute of Coastal Research, Swedish University of Agricultural Sciences, Öregrund, Sweden, **3** Department of Ecology, Environment and Plant Sciences, Stockholm University, Stockholm, Sweden, **4** Zoological Institute, University of Basel, Basel, Switzerland, **5** Department of Marine Sciences, University of Gothenburg, Gothenburg, Sweden

* ejakubaviciute@ekoi.lt

Citation: Jakubavičiūtė E, Bergström U, Eklöf JS, Haenel Q, Bourlat SJ (2017) DNA metabarcoding reveals diverse diet of the three-spined stickleback in a coastal ecosystem. PLoS ONE 12(10): e0186929. https://doi.org/10.1371/journal.pone.0186929

Editor: Mehrdad Hajibabaei, University of Guelph, CANADA

Received: March 10, 2017

Accepted: October 10, 2017

Published: October 23, 2017

Data Availability Statement: The raw data are available from the NCBI sequence read archive under accession number SRP101702, BioProject number PRJNA378633.

Funding: This study was funded by the Swedish Research Council for Environment, Agricultural Sciences and Spatial Planning (Formas, www.formas.se, grant 2013-1074), HM Carl XVI Gustaf's Foundation for Science and Education (2014-0002), the Baltic Sea Centre (Askö grants) and in-kind support from Groningen University (The

Abstract

The three-spined stickleback (*Gasterosteus aculeatus* L., hereafter 'stickleback') is a common mesopredatory fish in marine, coastal and freshwater areas. In large parts of the Baltic Sea, stickleback densities have increased >10-fold during the last decades, and it is now one of the dominating fish species both in terms of biomass and effects on lower trophic levels. Still, relatively little is known about its diet—knowledge which is essential to understand the increasing role sticklebacks play in the ecosystem. Fish diet analyses typically rely on visual identification of stomach contents, a labour-intensive method that is made difficult by prey digestion and requires expert taxonomic knowledge. However, advances in DNA-based metabarcoding methods promise a simultaneous identification of most prey items, even from semi-digested tissue. Here, we studied the diet of stickleback from the western Baltic Sea coast using both DNA metabarcoding and visual analysis of stomach contents. Using the cytochrome oxidase (CO1) marker we identified 120 prey taxa in the diet, belonging to 15 phyla, 83 genera and 84 species. Compared to previous studies, this is an unusually high prey diversity. Chironomids, cladocerans and harpacticoids were dominating prey items. Large sticklebacks were found to feed more on benthic prey, such as amphipods, gastropods and isopods. DNA metabarcoding gave much higher taxonomic resolution (median rank genus) than visual analysis (median rank order), and many taxa identified using barcoding could not have been identified visually. However, a few taxa identified by visual inspection were not revealed by barcoding. In summary, our results suggest that the three-spined stickleback feeds on a wide variety of both pelagic and benthic organisms, indicating that the strong increase in stickleback populations may affect many parts of the Baltic Sea coastal ecosystem.

Netherlands). This study was supported by Swedish research council grant 2012-3446 'Biodiversity genomics: Species identification pipelines for analysing marine invertebrate larval stages, community structure, and trophic interactions' (http://vrproj.vr.se/detail.asp?arendeid=92750) awarded to SJB. This study is a product of project Plant-Fish (www.plantfish.se).

Competing interests: The authors have declared that no competing interests exist.

Introduction

The three–spined stickleback (*Gasterosteus aculeatus* L., hereafter 'stickleback') is a common mesopredatory fish of high ecological interest, widespread all over the northern hemisphere in various habitats including coastal seas, estuaries, freshwater lakes and streams [1]. The stickleback is also an eco-genomic model organism, well-studied in terms of behavioural and evolutionary ecology [2–4]. Knowledge on the role of sticklebacks in aquatic food webs is, however, rather limited, especially in coastal and marine areas. To better understand the ecological role of sticklebacks, their feeding patterns and diet preferences need to be described, as feeding behaviour may affect community composition and food web functions.

In the brackish Baltic Sea, stickleback abundance has increased more than 10-fold during the last decade [5]. Currently, it constitutes up to 10% of the planktivorous biomass in offshore areas [5,6], and dominates fish assemblages in some coastal areas during summer, when adults immigrate from the open sea to spawn [6–8]. Experiments and field surveys indicate that sticklebacks may alter coastal food webs by feeding on and influencing lower trophic levels (e.g. grazers) [9], and worsening the effects of nutrient enrichment through cascading effects that increase the biomass of filamentous algae [6,7,10,11]. Moreover, sticklebacks may suppress populations of large predatory fish, such as northern pike and Eurasian perch, by predation on eggs and larvae, and the intraguild predation between sticklebacks and these large predatory fish may contribute to destabilizing food webs [5,12,13]. Thus, the increasing abundances of sticklebacks, in combination with their central role in ecosystem functioning, points to the need for more detailed knowledge on stickleback diets.

However, fish diet studies are challenging, with different methods having their own set of limitations. For the last century, the standard has been to visually identify prey from stomach contents, based on prey morphology [14]. This time-consuming method relies heavily on taxonomic expertise and can only be done when prey organisms are not too digested. Because most prey organisms rapidly degrade in stomachs, a high taxonomic resolution is often not possible, and a significant share of the prey tissue in the guts is often unidentifiable (e.g., [15]). A highly promising alternative to visual prey identification is metabarcoding methods, which combine DNA-based identification and high-throughput DNA sequencing, using taxonomically broad PCR primers to mass-amplify DNA barcodes from bulk samples (such as environmental samples or gut contents) [16]. Metabarcoding enables the identification of most prey items, even when diets are broad and diverse [17], and the simultaneous analysis of many samples. The aim of this study was to investigate the diet of the three-spined stickleback in coastal areas of the western central Baltic Sea, using a combination of classic (visual) and emerging (DNA metabarcoding) methods. Specifically, we addressed three questions: 1) what do sticklebacks eat in coastal areas, 2) how does stickleback diet depend on its body size, and 3) how do visual and DNA-based methods compare in terms of prey identification from stomach content. Accurate diet determination will provide more comprehensive information on coastal food webs, knowledge which is highly relevant in the context of ecosystem-based management to assess and potentially counteract the undesirable effects of massive increases of sticklebacks on the ecosystem [10,18].

Material and methods

This study was made in accordance with the ethical regulation laid down in the Swedish ordinance SJVFS 2012:26, which is the Swedish implementation of the Directive 2010/63/EU of the European Parliament and of the Council on the protection of animals used for scientific purposes. The fish died in the process of lifting the nets; after sticklebacks were removed from the nets they were immediately put in 95% ethanol. The fish sampling procedures applied in

the project were also judged and approved by the Ethical Board on Animal Experiments of the County Court of Uppsala, Sweden, permit C 139/13.

Study sites and sample collection

Sampling was performed in May 2014, after adult sticklebacks had migrated from their offshore winter areas into the coastal zone to spawn. Sampling was conducted in 16 bays situated along a 350 km stretch of the central Swedish Baltic Sea coast (Fig 1). Shallow bays are important reproduction areas for many coastal fish species, including sticklebacks [19]. They are characterized by a diverse and highly productive community of aquatic vegetation and macroinvertebrates, many of which constitute potential prey for sticklebacks [20]. The 16 bays were selected to represent a mix of shallow bay habitats along an archipelago gradient from the mainland to the outer archipelago, including sheltered shallow lagoons with narrow inlets, to more open and exposed bays.

Stickleback sampling. Sampling of stickleback stomachs was performed as part of a larger survey targeting the whole food webs of shallow vegetated bays (see [11]). Sticklebacks were caught using Nordic survey gillnets (European Union 112 standardized method EN 14757:2005). The nets were set at 0.5–3 m depth between 4–7 pm, and lifted between 7–9 am the following morning. The fish died in the process of lifting the nets; after sticklebacks were removed from the nets they were immediately put in 95% ethanol. In total, 196 individual fish were analysed (Fig 1). In bays where fewer than 15 sticklebacks were caught, all of the fish were analysed with respect to their diet composition. For bays where more individuals were caught, a subset representing the size distribution in the catch was selected for the diet analyses.

The total length (TL) of each fish was measured to the nearest 1 mm. The mean total length was 57.7 ±7.6 (SD), with a range of 35–72 mm (S1 Fig). Visual inspection of the resulting size frequency distribution indicates a left skew, i.e. an underrepresentation of large individuals (S1 Fig), which was not an effect of skewed subsampling. Only 2.5% (5 of the 196 individuals) were >70 mm; a much smaller proportion than that found in other, similar surveys in the Western Baltic Sea (unpublished; [5]).

Visual analysis of stomach content

Out of the 196 sticklebacks sampled, 192 were analysed using both visual methods and metabarcoding, and four were used in a pilot study for DNA metabarcoding. The stomachs were dissected and flushed with 80% EtOH to remove all stomach contents. To avoid cross-contamination, the dissection tools were rinsed with soap, bleach, and Milli-Q water before each individual dissection. Prey items visually distinguishable in the flushed stomach contents were identified to the highest taxonomic resolution possible, using a stereo microscope (magnification 20-80x). Frequency of occurrence for each prey item was estimated ($\%F_{vis}$, the percentage of stomachs in which a prey was present). Thereafter, all stomach contents were stored at -20˚C in 80% EtOH for subsequent DNA extraction. The level of digestion for each stomach was classified on a 1–5 scale, where 1 = intact prey, 2 = partially digested, 3 = extensively digested, 4 = very few prey parts discernible, and 5 = fully digested/ empty stomach.

DNA metabarcoding analysis

Sample processing. DNA was extracted from the 196 sticklebacks' gut contents using the UltraClean® Tissue and Cells DNA Isolation Kit (MO BIO Laboratories), according to the manufacturer's instructions. The dual PCR amplification method was used for Illumina MiSeq library preparation [21]. The cytochrome oxidase 1 (CO1) marker was first amplified using

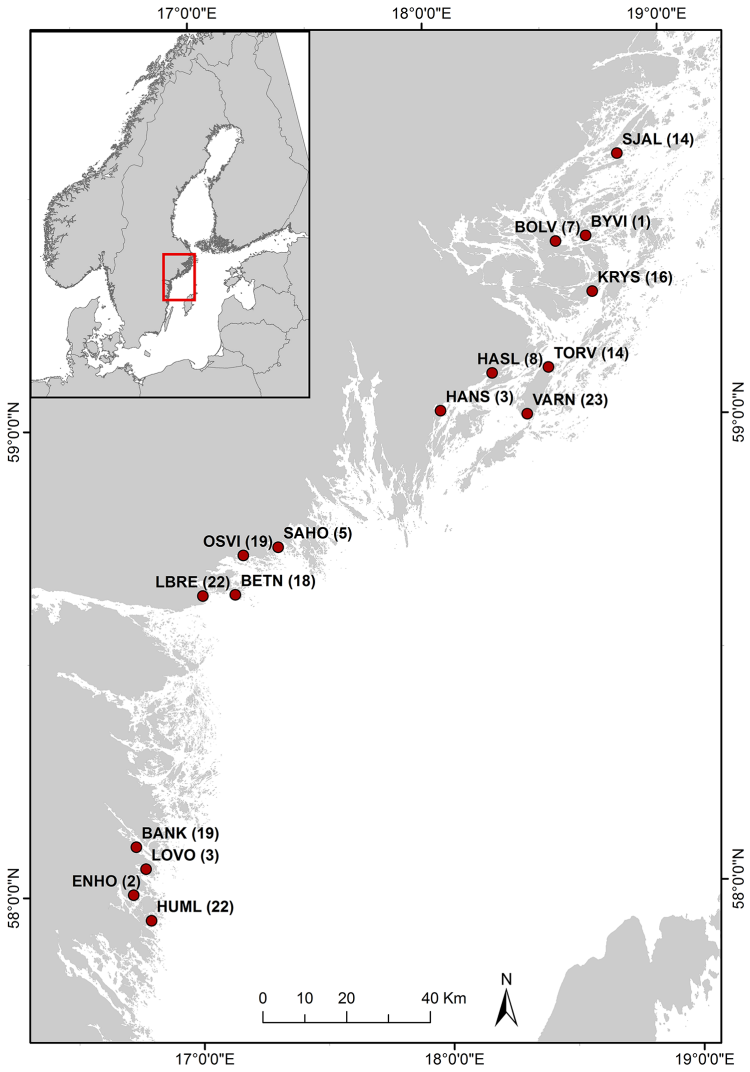


Fig 1. Sampling sites. Numbers in brackets indicate number of sticklebacks analysed from each bay.

https://doi.org/10.1371/journal.pone.0186929.g001

locus specific primers including an Illumina adapter overhang (amplicon PCR). The primers were based on Leray et al.'s (2013) [22] 'mini-barcode' yielding a 313 bp fragment (CO1mini_mI COIintF_MiSeq: TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGGWACWGGWTGAACWGTWT AYCCYCC, CO1_dgHCO2198_MiSeq: GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTA AACTTCAGGGTGACCAAARAAYCA, CO1 specific sequence is shown in bold, and illumina adapter in regular font). A blocking primer was used in the amplicon PCR, to prevent amplification from

Table 1. Number of reads after each bioinformatic data processing step.

Paired-end joining	Primer trimming	Quality filtering
15 706 724	10 982 728	10 586 546

https://doi.org/10.1371/journal.pone.0186929.t001

G. aculeatus, following [23]. A Spacer C3 CPG was added to the 3' end of the blocking primer to prevent elongation without affecting annealing properties, minimizing predator DNA amplification (G_aculeatus_block_Hco_2198: CAAAGAATCAAAATAAGTGTTGGTAAAGA-C3). For each sample, two independent PCR reactions were performed and later pooled, ensuring greater coverage of prey items amplified. In a second PCR step, Illumina dual index adapters were incorporated to the amplicons using a limited number of cycles (Index PCR).

Amplicon PCRs were performed as 30 μl reactions with 20pm of each primer and 100pm of blocking primer and using Pfu proofreading DNA polymerase (Promega). Cycling conditions were as follows: 2 min at 95˚C (1x); 1 min at 95˚C, 45s at 55˚C, 1 min at 72˚C (40x); 5 min at 72˚C (1x); hold at 4˚C. Amplicons were checked on a 2% agarose gel. Agencourt® AMPure® XP paramagnetic beads (Beckman Coulter) were then used to purify the PCR products [21]. For index PCR, the Illumina Nextera XT kit (96 indices, 384 samples) was used according to manufacturer's instructions. Index PCR was performed as 50 μl reactions using 5 μl of cleaned up amplicons. Cycling conditions were as follows: 3 min at 95˚C (1x); 30s at 95˚C, 30s at 55˚C, 30s at 72˚C (8x); 5 min at 72˚C (1x); hold at 4˚C. Agencourt® AMPure® XP paramagnetic beads (Beckman Coulter) were then used to purify the PCR products, using a ratio of 0.8 that allows the selection of fragments larger than 200 bp. DNA quantification was carried out using a Qubit Fluorometer (Invitrogen) and the average fragment size was verified using Tapestation (Agilent Technologies). Pooled libraries were then sequenced as paired-ends using Illumina MiSeq Reagent v3, producing 30 103 790 paired-end reads of 300 bp in length.

Bioinformatic data processing and analysis. The processing steps were performed using Qiime (Quantitative Insights into Microbial Ecology) version 1.9.1 [24] and custom python scripts. Paired-end joining was done using the Qiime fastq-join tool. A 48% sequence loss was observed after the paired-end joining step due to poor sequence quality at read ends (the raw data are available from the NCBI sequence read archive under accession number SRP101702, BioProject number PRJNA378633). Dual indexes and Illumina overhangs were removed by the sequencing platform. Primer sequences were removed using a custom python script (https://github.com/Quiterie90/Primer_Removal), corresponding to a 30% loss (Table 1). Due to its stringency, the script quality filters sequences by removing incomplete reads or chimeras. Additional quality filtering with Qiime removed 3% of the reads. Finally, remaining chimeric reads were excluded using UCHIME [25], producing a final dataset (0.5% loss).

The Bayesian clustering algorithm CROP was used to cluster the sequences into operational taxonomic units (OTUs) based on the natural distribution of the data, using a Gaussian model [26]. According to a benchmarking study by Leray et al. [22], the best lower and upper bound values to cluster metazoan CO1 sequences are 3 and 4, corresponding to sequence dissimilarities between 6% and 8% (CROP -i <input.fasta> -b 211731 -z 470 -l 3 -u 4 -o <output>).

For taxonomic assignment of CO1 sequences, a custom database was created, consisting in a taxonomy file associated with a reference sequence file, of Metazoan sequences retrieved from BOLD (http://www.boldsystems.org/ downloaded in March 2016), combined with own reference databases of Chironomidae, Nemertea, Xenacoelomorpha and Oligochaeta and barcodes of Swedish Echinodermata, Mollusca, Cnidaria and Arthropoda from the Swedish Barcode of Life database (SweBol).

Taxonomic assignment was done using a 97% similarity threshold using the Uclust software implemented in Qiime with the default parameters [27]. In order to obtain matches for non-Metazoan taxa, we also did a Megablast search with a 97% similarity threshold, a minimum query coverage of 70% and an e-value inferior to 1E-100 against the Genbank nt (nucleotide) database (ftp://ftp.ncbi.nlm.nih.gov/blast/db/) with Geneious [28].

Data analysis. After sequencing, we obtained an OTU table showing the number of reads per taxon found in the stomach of each fish. For diet derived from this barcoding identification, frequency of occurrence was estimated ($\%F_{bar}$)—the percentage of stomachs in which a prey (OTU) was present.

To investigate the effect of fish body size (mm TL) on their diet and account for the hierarchical data structure, we performed permutational multivariate analysis of variance (PERMANOVA, *adonis* function in the vegan package for R [29] on the Bray–Curtis distance matrix with 'bay' (16 levels) as strata, fish size group as fixed predictor, and diet composition (counts of stomach with a certain prey present) as a response. Fish were divided into two size groups (TL): $\leq$6.5 cm (S), and >6.5 cm (L).

Comparison of visual vs DNA-based methods

The results from the visual analysis and the metabarcoding analysis were compared with respect to both number of taxa identified and to the taxonomic resolution of the data. The number of taxa was the mean number of taxa identified per stickleback in the two methods applied. To compare the methods with respect to their taxonomic resolution, ranks were given to each prey item in each stomach and then mean taxonomic rank of the stomach was used [30]. Taxonomic resolution was ranked as follows: species = 1, genus = 2, family = 3, infra-order = 4, order = 5, infra-class = 6, class = 7, phylum = 8. Infra-class and infra-order represent taxonomic rankings between class and order and between family and order, respectively. Paired t-tests were used to compare the resolution between the methods.

Results

Diet composition based on DNA barcoding

Using metabarcoding, 120 taxa were identified in the stomachs of sticklebacks: 15 phyla, 27 classes, 52 orders, 66 families, 83 genera, and 84 species (S1 Table). A broad range of phyla were found, but Arthropoda dominated by far (Fig 2). Given that this is the first barcoding-based study of Baltic Sea stickleback diet, we provide the whole list of taxa found (S1 Table). We only omit records from primates and birds, which were obviously contamination. Taxa likely to be accidental or secondary prey were also excluded from further analyses. Specifically, we excluded Fungi, Macroalgae and Chromista (as these are not targeted as food by sticklebacks), and kept only Metazoa in the primary prey list. A few OTUs of Metazoa were also excluded as they were either unlikely to be prey, or due to possible contamination (see S1 Table). In total, 103 taxa were considered primary prey and were used in the subsequent analyses.

Sticklebacks had a broad spectrum of prey items, of which Insecta (mainly chironomids), Maxillipoda (harpacticoid copepods) and Branchiopoda (cladocerans) were the dominating food items, found in more than 90% of the samples (S1 Table). At the species level, the main prey were the chironomid *Tanytarsus usmaensis*, the harpacticoid *Tachidius discipes*, and the cladoceran *Pleopis polyphemoides* (S1 Table).

Although the range of stickleback body lengths was too small to detect ontogenetic diet shifts, significant differences in stomach content depending on fish size were found (PERMANOVA, $F = 3.7$, $p = 0.01$). The diet of the large fish (>6.5 cm) differed from the group of

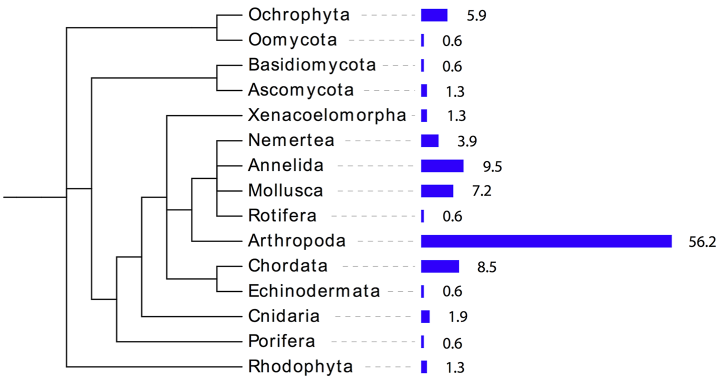


Fig 2. Frequencies of phyla identified in three-spined stickleback gut contents. Bar length corresponds to the frequency of OTU assigned to a specific phylum.

https://doi.org/10.1371/journal.pone.0186929.g002

smaller fish (≤6.5 cm). Specifically, amphipods, isopods and gastropods appeared to be more common in the diet of the larger fish, as well as insects like hemipterans and coleopterans (Fig 3).

Methods comparison: Visual identification vs DNA barcoding

The taxonomic resolution of the prey identified differed substantially between the two methods. DNA barcoding gave a much higher resolution (with median rank of genus, $p<0.0001$).

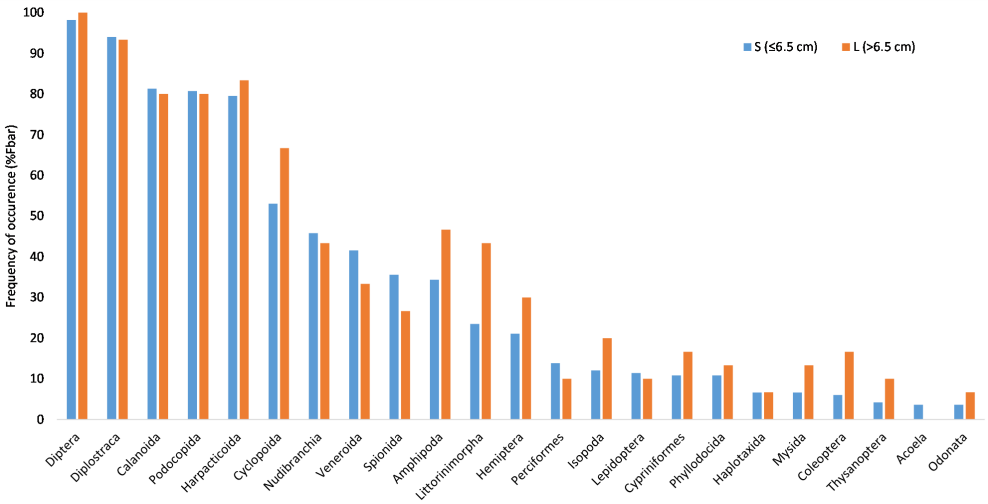


Fig 3. Diet composition of different size groups of stickleback (at order level). Only orders with >5% of frequency of occurrence are shown.

https://doi.org/10.1371/journal.pone.0186929.g003

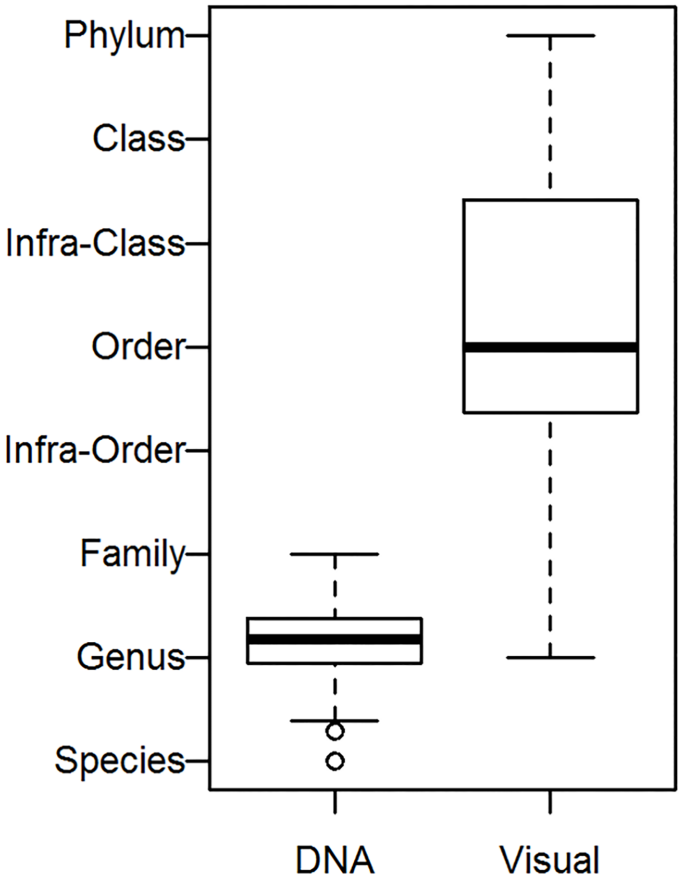


Fig 4. Mean taxonomic rank assigned to items within individual stomachs. DNA–assigned by barcoding, Visual–visual identification disregarding non-identified items. Midline represents median, boxes first and third quartiles, whiskers either maximum values or 1.5 times interquartile range (whichever is smaller) and circles outliers.

https://doi.org/10.1371/journal.pone.0186929.g004

Disregarding stomachs for which no visual identification could be done, the median taxonomic rank for visual inspection was order (Fig 4).

DNA barcoding also resulted in a much higher number of prey taxa identified per stomach than visual analysis ($p<0.0001$): 21.7±8.8 vs 1.96±1 (mean±SD). Also the total number of taxa identified using DNA barcodes was much larger than the number of taxa identified using visual quantification (120 vs 21; see S1 and S2 Tables). The average level of digestion was 3.6, meaning that gut contents were extensively digested and/or with very few prey items present. Not surprisingly, many taxa identified using DNA barcodes could not have been identified visually (e.g. due to their small size). However, some taxa identified by visual inspection were not revealed by barcoding (*Temora longicornis*, *Bosminidae*, *Hydracarina*).

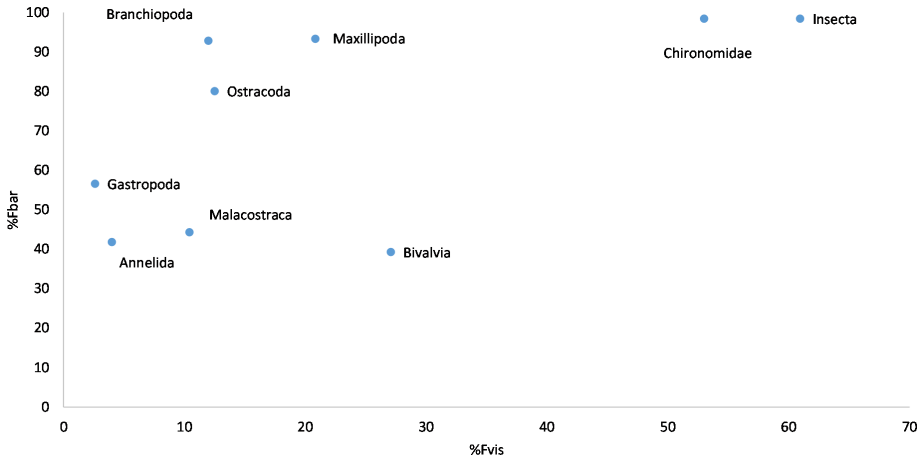


Fig 5. Diet of three-spined stickleback. Relationship between the results of two methods used: frequency of occurrence determined by metabarcoding ($\%F_{bar}$) and by visual analysis ($\%F_{vis}$).

https://doi.org/10.1371/journal.pone.0186929.g005

Irrespective of these minor differences, the two methods showed consistent patterns: at the population level, frequency of occurrence determined by visual analysis ($\%F_{vis}$) corresponded well with the frequencies based on OTU reads ($\%F_{bar}$; Fig 5), although the relationship was not linear. Instead, a curvilinear relationship was seen, where the frequencies of occurrence in the metabarcoding analyses were higher than in the visual identification for all taxa. Such a relationship is to be expected, since barcoding is capable of detecting even very little amounts of the prey, which could not be detected visually. Only Bivalvia appeared to have very similar frequencies detected by both methods.

Discussion

The aim of this study was to assess the diet of the three-spined stickleback in a coastal ecosystem, and to compare classic visual analysis and novel DNA metabarcoding for identifying fish prey in stomach contents. The main prey items found were chironomids, cladocerans and harpacticoid copepods. Large (>6.5 cm) sticklebacks had higher proportions of benthic herbivores, like amphipods, gastropods and isopods, in their diet. The results of the DNA barcoding revealed a highly diverse stickleback diet (more than 100 taxa in total, and >20 per individual) and provided a much higher taxonomic resolution than the conventional visual stomach content analysis.

Diet composition

While stickleback diets are well studied from other parts of the world, no previous studies have revealed such a high diversity of prey items, most likely because of limitations in the methods used (for some examples of previous studies, see S3 Table). Sticklebacks, however, inhabit many different habitats and ecosystems, so their diet varies accordingly. In pelagic areas of the Baltic Sea, where sticklebacks spend a large part of their life, they feed primarily on cladocerans and calanoid copepods [31–33], but at the coast the main prey items are insect larvae, harpacticoids and amphipods [34,35]. In freshwater systems, they are known to prey on both planktonic and

benthic prey. We found chironomids and harpacticoids to be a very important part of the stickleback diet, similar to the diet in the coastal zone of the Bothnian Bay, in the northern Baltic Sea [35].

From a food-web perspective, the high abundance of cladocerans found in the diet (Fig 3, S1 Table) might indicate competition with juvenile stages of other fish, especially when preference for cladocerans is evident [33,36,37]. Ljunggren et al. (2010) [6] suggested that recruitment of coastal predatory fish in the Baltic Sea (pike and perch) was impaired by limited food availability (zooplankton) for their larvae, due to competition with sticklebacks. The three-spined stickleback has indeed been shown to deplete zooplankton communities in brackish water lagoons with similar densities as in the current study area [38]. On the other hand, sticklebacks have also been shown to feed on small pike and perch larvae, which would constitute a more direct effect on populations of large predators, than competition [12]. We detected *Perciformes* in the stomachs of six fish (see S1 Table), potentially indicating sticklebacks may have been feeding on perch egg or larvae.

Concerning benthic prey, the most significant part of the diet consisted of chironomid larvae, which were one of the most common epibenthic organism groups in the 16 bays. Sticklebacks are well-known to feed on chironomids in freshwater areas ([35,36,39,40], S3 Table). Chironomids are a broad taxonomic group, with a diverse diet spanning between phytoplankton, epiphytic algae, detritus, macrophytes, and crustacean zooplankton [41]. More knowledge on the role of chironomids in the food webs and the interactions with sticklebacks is needed, since possible cascading effects from sticklebacks via chironomids to lower trophic levels may be present (e.g., [42]).

Sticklebacks seemed to have fed less on the gammarid amphipods than expected from previous experimental studies, where they have been shown to strongly reduce gammarid densities in the lab and in the field [7,9]. In our study, larger sticklebacks appear to feed more on amphipods (Fig 3). It is well known that stickleback mouth width and gape size influence the size of prey that can be eaten [43,44], and that jaw morphology (gape size, gill raker spacing) can change food handling efficiency [45]. Therefore, the optimum diet might differ between stickleback populations and/or habitats depending on their morphology. Hart and Ison (1991) found the size threshold of prey rejection to be at 6–7 mm [44], Byström et al. (2015) suggests the upper limit to be around 5 mm [12]. Given that fish above 5 cm eat amphipods [34,46], there were plenty of gammarids of appropriate size for sticklebacks to eat (see S2 Fig), showing that mouth morphology does not explain rejection of amphipods in small stickleback.

The most likely explanation for the relatively low proportion of gammarids in the overall diet is an underrepresentation of large individuals sampled in this study (S1 Fig) compared to several previous studies ([5]; and unpublished). These large individuals appear to feed the most on gammarids (Fig 3). The underrepresentation in the nets, which were placed at $>$ 1m depth, may indicate that the largest sticklebacks occupy the most beneficial habitats in the bays, i.e. the shallowest vegetated parts, where we could not fish using gillnets. These shallow areas are also the habitats with the highest abundances of gammarids, which may have led to the low frequency of stickleback predation upon gammarids apparent in the analysis. Large sticklebacks (>6.5 cm) also tend to have a higher frequency of occurrences of cyclopoid copepods than smaller ones (Fig 3), mainly driven by *Eucyclops macruroides*. This species is known to inhabit vegetation in the littoral zone, which again supports a possible small-scale differences in foraging habitats between stickleback size classes.

In many of the 16 bays there were relatively few stickleback individuals sampled, resulting in the inability to assess individual specialisation. To assess the link between sticklebacks and large, benthic crustaceans (e.g., gammarids), more detailed and intense sampling should be conducted, and the potential for individual specialisation should be investigated.

Visual inspection vs DNA barcoding

In general, the two methods gave consistent results with the same prey taxa dominating (Fig 5). However, as the stomach content was extensively digested and/or with very few prey items present, the visual prey species identification was in many cases obscured. Fish may have been caught in the nets up to 12 hours before they were preserved, making the visual analysis particularly difficult. On the other hand, this may also have had an effect on DNA degradation. In diet studies, a high proportion of unidentifiable material in the guts, which cannot be visually assigned to any prey category, is common [15]. Even though both methods are time-consuming and expensive, and despite the fact that some prey species were missed (*Temora longicornis*, *Bosminidae*, Hydracarina), barcoding provided a much higher taxonomic resolution and therefore produced a more accurate and detailed analysis of gut contents In terms of the resolution provided by the two methods, our results are similar to a previous fish feeding ecology study [30].Thus, we consider these discrepancies as minor, since barcoding still enabled the disclosure of unexpectedly high diversity in the stickleback diet.

Methodological shortcomings

The results of DNA metabarcoding and the visual analysis did not match fully.—Some prey taxa (*Temora longicornis*, *Bosminidae*, Hydracarina) were detected by visual inspection only, while Bivalvia had very similar frequency values estimated by both methods (see Fig 5). As we could visually identify these prey organisms, their DNA is unlikely to have been too degraded for barcoding to identify them. A more likely explanation is that even though the CO1 primers are designed to be taxonomically broad they may not bind equally well to all prey species, and maybe not at all to some. It is known that even minor primer–template mismatches can lead to substantial under-representation of the prey in the diet [47]. These biases are then accumulated through DNA amplifications during the PCR reaction [48,49]. Bosminidae was identified during visual inspection of stomach contents, but when barcoded only a higher corresponding taxon was detected (Diplostraca). Thus, only species or group specific primers would guarantee the most accurate identification.

Blocking primers are used to avoid 'predator sequences' (i.e. lots of non-informative reads), which can reduce potential of prey detection [50], but could also block prey DNA [51], which may bring in bias when analysing mixtures of DNA. We used a blocking primer to avoid stickleback sequences, but since predator and prey missed are not phylogenetically close, and the blocking primer used is specific to *G. aculeatus*, this should not have impaired the results.

There can be other biases introduced during the bioinformatic analysis steps, such as during the clustering of sequences, where the number of OTUs or 'species' found depends on the sequence similarity cut-off used, and during taxonomic assignment, which uses a sequence identity threshold of 97%. Also, it is obvious that if some species are not represented in the DNA reference library, no matches for these will be found.

Secondary consumption, i.e., prey of the prey, parasites or accidental material consumed during feeding, may confound the results in DNA-based studies [52–54]. The magnitude of potential error due to secondary predation depends on digestion rates [54]. We acknowledge that even though a few unlikely prey taxa were removed from the analysis, some secondary prey may still have been included in the analysis as primary prey items. However, DNA of secondary prey might be expected to represent only a minor part of total OTU reads compared to primary prey, due to a much lower total biomass and to a higher level of degradation.

When visually inspecting the often highly degraded stomach content, prey items such as fish eggs and larvae may be substantially underestimated (e.g., [55]). Although metabarcoding has the power to catch such prey species, the life stages of prey items remain unknown.

Moreover, prey analyses based on stomach content only represent a snapshot in time. To obtain more comprehensive knowledge on stickleback diets, future studies should be complemented with analyses of stable isotopes/fatty acids, which integrate the signal from different prey organisms over longer time.

Despite these shortcomings, DNA metabarcoding seems to be a viable method to assess stickleback diets. From a data quality perspective, we therefore, at least until metabarcoding methods are further developed, suggest to combine high-throughput DNA sequencing and traditional visual stomach content analysis, to achieve the best resolution of diet composition and diversity.

Implications

Using a powerful combination of visual and metabarcoding-based analyses of stomach contents, we show that the three-spined stickleback feeds on a wide variety of organisms in coastal areas of the Baltic Sea, including pelagic zooplankton and benthic epifaunal invertebrates. As a consequence, the major increase in stickleback abundance [5] could affect many parts of both pelagic and benthic food webs, resulting in competition with other fish species, and cascading effects down to primary producers [7,11]. Given that the expected increase in the Baltic Sea surface water temperatures [56] may be beneficial for stickleback population growth [57], studies such as this one could provide important information about the current and future impacts of three-spined sticklebacks on the Baltic Sea ecosystem.

Supporting information

S1 Table. Taxa found in three-spined stickleback stomachs as revealed by DNA metabarcoding (Primates and Aves excluded). Items in italics were considered as secondary/ accidental prey $\%F_{bar}$- frequency of occurrence (percentage of stomachs in which a prey was present).
(DOCX)

S2 Table. Diet of three-spined stickleback as revealed by visual stomach content analysis. $\%F_{vis}$—the percentage of stomachs in which a prey was present.
(DOCX)

S3 Table. Summary of some studies on three-spined stickleback diet.
(DOCX)

S1 Fig. Stickleback size (total length, mm) distribution in a sample.
(TIF)

S2 Fig. Gammaridae size distribution in the bays studied.
(TIF)

S1 Appendix. Comparison of quantification from OTU reads and results of visual stomach content analysis.
(DOCX)

Acknowledgments

We thank Å Nilsson Austin J Hansen, S Donadi, BK Eriksson, P Jacobsson, G Johansson, M van Regteren, S Skoglund, M van der Snook, G Sundblad and V Thunell for field and lab assistance; Jennie Finnman at the Genomics Core facility platform, Sahlgrenska Academy, University of Gothenburg for help with Illumina MiSeq. We thank Matthieu Leray for sharing his knowledge and for help with blocking primer design. We also thank the SweBoL (Swedish

Barcode of Life) network for sharing unpublished sequence data on Swedish invertebrates, Christer Erséus and Per Sundberg for sharing barcode databases of Nemerteans and Oligochaetes, and especially Thomas Lyrholm for sharing an unpublished database of chironomid sequences.

Author Contributions

Conceptualization: Ulf Bergström, Johan S. Eklöf.

Data curation: Quiterie Haenel, Sarah J. Bourlat.

Formal analysis: Eglė Jakubavičiūtė, Quiterie Haenel.

Funding acquisition: Ulf Bergström, Johan S. Eklöf.

Investigation: Eglė Jakubavičiūtė, Ulf Bergström, Johan S. Eklöf, Quiterie Haenel.

Methodology: Eglė Jakubavičiūtė, Ulf Bergström, Johan S. Eklöf, Quiterie Haenel, Sarah J. Bourlat.

Project administration: Eglė Jakubavičiūtė, Ulf Bergström, Johan S. Eklöf, Sarah J. Bourlat.

Resources: Johan S. Eklöf, Sarah J. Bourlat.

Software: Quiterie Haenel.

Supervision: Ulf Bergström, Johan S. Eklöf, Sarah J. Bourlat.

Validation: Eglė Jakubavičiūtė, Sarah J. Bourlat.

Visualization: Eglė Jakubavičiūtė, Quiterie Haenel.

Writing – original draft: Eglė Jakubavičiūtė, Johan S. Eklöf.

Writing – review & editing: Eglė Jakubavičiūtė, Ulf Bergström, Johan S. Eklöf, Quiterie Haenel, Sarah J. Bourlat.

References

1. Bell MA, Foster SA. The evolutionary biology of the threespine stickleback. Oxford: Oxford University Press. 1994.
2. Huntingford FA, Ruiz-Gomez ML. Three-spined sticklebacks Gasterosteus aculeatus as a model for exploring behavioural biology. J Fish Biol. 2009; 75: 1943–1976. https://doi.org/10.1111/j.1095-8649.2009.02420.x PMID: 20738667
3. Hendry AP, Peichel CL, Matthews B, Boughman JW, Nosil P. Stickleback research: The now and the next. Evol Ecol Res. 2013; 15: 111–141.
4. Des Roches S, Shurin JB, Schluter D, Harmon LJ. Ecological and evolutionary effects of stickleback on community structure. PLoS One. 2013; 8: e59644. https://doi.org/10.1371/journal.pone.0059644 PMID: 23573203
5. Bergström U, Olsson J, Casini M, Eriksson BK, Fredriksson R, Wennhage H, et al. Stickleback increase in the Baltic Sea–A thorny issue for coastal predatory fish. Estuar Coast Shelf Sci. 2015; 163: 1–9. https://doi.org/10.1016/j.ecss.2015.06.017
6. Ljunggren L, Sandstrom A, Bergström U, Mattila J, Lappalainen A, Johansson G, et al. Recruitment failure of coastal predatory fish in the Baltic Sea coincident with an offshore ecosystem regime shift. ICES J Mar Sci. 2010; 67: 1587–1595. https://doi.org/10.1093/icesjms/fsq109
7. Sieben K, Ljunggren L, Bergström U, Eriksson BK. A meso-predator release of stickleback promotes recruitment of macroalgae in the Baltic Sea. J Exp Mar Bio Ecol. Elsevier B.V.; 2011; 397: 79–84. https://doi.org/10.1016/j.jembe.2010.11.020
8. Park PJ, Aguirre WE, Spikes DA, Miyazaki JM. Landmark-Based Geometric Morphometrics: What Fish Shapes Can Tell Us about Fish Evolution. Proc Assoc Biol Lab Educ. 2013; 34: 361–371.

9. Sieben K, Rippen AD, Eriksson BK. Cascading effects from predator removal depend on resource availability in a benthic food web. Mar Biol. 2011; 158: 391–400. https://doi.org/10.1007/s00227-010-1567-5 PMID: 24391255
10. Eriksson BK, Sieben K, Eklöf J, Ljunggren L, Olsson J, Casini M, et al. Effects of altered offshore food webs on coastal ecosystems emphasize the need for cross-ecosystem management. Ambio. 2011; 40: 786–797. https://doi.org/10.1007/s13280-011-0158-0 PMID: 22338716
11. Donadi S, Austin ÅN, Bergström U, Eriksson BK, Hansen JP, Jacobson P, et al. A cross-scale trophic cascade from large predatory fish to algae in coastal ecosystems. Proc R Soc B Biol Sci. 2017; 284.
12. Byström P, Bergström U, Hjälten A, Ståhl S, Jonsson D, Olsson J. Declining coastal piscivore populations in the Baltic Sea: Where and when do sticklebacks matter? Ambio. 2015; 44: 462–471. https://doi.org/10.1007/s13280-015-0665-5 PMID: 26022328
13. Nilsson J. Predation of Northern Pike (Esox lucius L.) Eggs: A Possible Cause of Regionally Poor Recruitment in the Baltic Sea. Hydrobiologia. 2006; 553: 161–169. https://doi.org/10.1007/s10750-005-1949-8
14. Hyslop EJ. Stomach contents analysis-a review of methods and their application. J Fish Biol. 1980; 17: 411–429. https://doi.org/10.1111/j.1095-8649.1980.tb02775.x
15. Baker R, Buckland A, Sheaves M. Fish gut content analysis: robust measures of diet composition. Fish Fish. 2014; 15: 170–177. https://doi.org/10.1111/faf.12026
16. Taberlet P, Coissac E, Pompanon F, Brochmann C, Willerslev E. Towards next-generation biodiversity assessment using DNA metabarcoding. Mol Ecol. 2012; 21: 2045–2050. https://doi.org/10.1111/j.1365-294X.2012.05470.x PMID: 22486824
17. Pompanon F, Deagle BE, Symondson WOC, Brown DS, Jarman SN, Taberlet P. Who is eating what: Diet assessment using next generation sequencing. Mol Ecol. 2012; 21: 1931–1950. https://doi.org/10.1111/j.1365-294X.2011.05403.x PMID: 22171763
18. Östman Ö, Eklöf J, Eriksson BK, Olsson J, Moksnes PO, Bergström U, et al. Top-down control as important as nutrient enrichment for eutrophication effects in North Atlantic coastal ecosystems. J Appl Ecol. 2016; 53: 1138–1147. https://doi.org/10.1111/1365-2664.12654
19. Snickars M, Sandstrom A, Lappalainen A, Mattila J, Rosqvist K, Urho L. Fish assemblages in coastal lagoons in land-uplift succession: The relative importance of local and regional environmental gradients. Estuar Coast Shelf Sci. 2009; 81: 247–256. https://doi.org/10.1016/j.ecss.2008.10.021
20. Hansen JP, Wikström SA, Kautsky L. Effects of water exchange and vegetation on the macroinvertebrate fauna composition of shallow land-uplift bays in the Baltic Sea. Estuar Coast Shelf Sci. 2008; 77: 535–547. https://doi.org/10.1016/j.ecss.2007.10.013
21. Begg G, Waldman J. An holistic approach to fish stock identification. Fish Res. 1999; 43: 35–44.
22. Leray M, Yang JY, Meyer CP, Mills SC, Agudelo N, Ranwez V, et al. A new versatile primer set targeting a short fragment of the mitochondrial COI region for metabarcoding metazoan diversity: application for characterizing coral reef fish gut contents. Front Zool. 2013; 10: 34. https://doi.org/10.1186/1742-9994-10-34 PMID: 23767809
23. Leray M, Agudelo N, Mills SC, Meyer CP. Effectiveness of Annealing Blocking Primers versus Restriction Enzymes for Characterization of Generalist Diets: Unexpected Prey Revealed in the Gut Contents of Two Coral Reef Fish Species. PLoS One. 2013; 8. https://doi.org/10.1371/journal.pone.0058076 PMID: 23579925
24. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, et al. QIIME allows analysis of high-throughput community sequencing data. Nat Methods. 2010; 7: 335–6. https://doi.org/10.1038/nmeth.f.303 PMID: 20383131
25. Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R. UCHIME improves sensitivity and speed of chimera detection. Bioinformatics. 2011; 27: 2194–2200. https://doi.org/10.1093/bioinformatics/btr381 PMID: 21700674
26. Hao X, Jiang R, Chen T. Clustering 16S rRNA for OTU prediction: A method of unsupervised Bayesian clustering. Bioinformatics. 2011; 27: 611–618. https://doi.org/10.1093/bioinformatics/btq725 PMID: 21233169
27. Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinformatics. 2010; 26: 2460–2461. https://doi.org/10.1093/bioinformatics/btq461 PMID: 20709691
28. Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, et al. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics. 2012; 28: 1647–9. https://doi.org/10.1093/bioinformatics/bts199 PMID: 22543367
29. Bolnick DI, Snowberg LK, Hirsch PE, Lauber CL, Knight R, Caporaso JG, et al. Individuals' diet diversity influences gut microbial diversity in two freshwater fish (threespine stickleback and Eurasian perch). Ecol Lett. 2014; 17: 979–987. https://doi.org/10.1111/ele.12301 PMID: 24847735

30. Berry O, Bulman C, Bunce M, Coghlan M, Murray DC, Ward RD. Comparison of morphological and DNA metabarcoding analyses of diets in exploited marine fishes. Mar Ecol Prog Ser. 2015; 540: 167–181. https://doi.org/10.3354/meps11524

31. Jakubavičiūtė E, Casini M, Ložys L, Olsson J. Seasonal dynamics in the diet of pelagic fish species in the southwest Baltic Proper. ICES J Mar Sci J du Cons. 2017; 74: 750–758. https://doi.org/10.1093/icesjms/fsw224

32. Peltonen H, Vinni M, Lappalainen A, Ponni J. Spatial feeding patterns of herring (L.), sprat (L.), and the three-spined stickleback (L.) in the Gulf of Finland, Baltic Sea. ICES J Mar Sci. 2004; 61: 966–971. https://doi.org/10.1016/j.icesjms.2004.06.008

33. Lankov A, Ojaveer H, Simm M, Põllupüü M, Möllmann C. Feeding ecology of pelagic fish species in the Gulf of Riga (Baltic Sea): the importance of changes in the zooplankton community. J Fish Biol. 2010; 77: 2268–84. https://doi.org/10.1111/j.1095-8649.2010.02805.x PMID: 21155782

34. Candolin U, Johanson A, Budria A. The Influence of Stickleback on the Accumulation of Primary Production: a Comparison of Field and Experimental Data. Estuaries and Coasts. 2016; 39: 248–257. https://doi.org/10.1007/s12237-015-9984-9

35. Frande C, Kjellman J, Leskela A, Hudd R. The food of three-spined stickleback (*Gasterosteus aculeatus*) on a whitefish (Coregonus lavaretus) nursery area in the bay of Bothnia. Aqua Fenn. 1993; 85–87.

36. Campbell CE. Prey Selectivities of Threespine Sticklebacks (*Gasterosteus aculeatus*) and Phantom Midge Larvae (*Chaoborus* Spp) in Newfoundland Lakes. Freshw Biol. 1991; 25: 155–167. https://doi.org/10.1111/j.1365-2427.1991.tb00481.x

37. Leinikki J. The diet of three-spined stickleback in the Gulf of Bothnia during its open water phase. Aqua Fenn. 1995; 25: 71–75.

38. Jakobsen TS, Hansen PB, Jeppesen E, Grønkjær P, Søndergaard M. Impact of three-spined stickleback *Gasterosteus aculeatus* on zooplankton and chl a in shallow, eutrophic, brackish lakes. Mar Ecol Prog Ser. 2003; 277–284. https://doi.org/10.3354/meps262277

39. Hynes H. The food of freshwater sticklebacks (*Gasterosteus aculeatus and Pygosteus pungitius*), with a review of methods used in studies of the food of fishes. J Anim Ecol. 1950; 19: 36–58. https://doi.org/10.2307/1570

40. Dukowska M, Grzybkowska M, Marszał L, Zięba G. The food preferences of three-spined stickleback, *Gasterosteus aculeatus* L., downstream from a dam reservoir. Oceanol Hydrobiol Stud. 2009; 38: 39–50. https://doi.org/10.2478/v10009-009-0020-x

41. Armitage PD, Pinder LC, Cranston P. The Chironomidae: Biology and ecology of non-biting midges. Springer Netherlands; 2012.

42. Rudman SM, Rodriguez-cabal MA, Stier A, Sato T, Heavyside J, El-sabaawi RW, et al. Adaptive genetic variation mediates bottom-up and top-down control in an aquatic ecosystem. Proc R Soc B Biol Sci. 2015; 282: 20151234. https://doi.org/10.1098/rspb.2015.1234 PMID: 26203004

43. Lavin PA, McPhail JD. Adaptive Divergence of Trophic Phenotype among Freshwater Populations of the Threespine Stickleback (*Gasterosteus aculeatus*). Can J Fish Aquat Sci. NRC Research Press; 1986; 43: 2455–2463. https://doi.org/10.1139/f86-305

44. Hart PJB, Ison S. The influence of prey size and abundance, and individual phenotype on prey choice by the 3-spined stickleback, *Gasterosteus aculeatus* l. J Fish Biol. 1991; 38: 359–372. https://doi.org/10.1111/j.1095-8649.1991.tb03126.x

45. Ibrahim AA, Huntingford FA. Foraging efficiency in relation to within-species variation in morphology in threeispined sticklebacks, *Gasterosteus aculeatus*. J Fish Biol. 1988; 33: 823–824. https://doi.org/10.1111/j.1095-8649.1988.tb05528.x

46. Kotta J, Orav-Kotta H, Herkül K. Separate and combined effects of habitat-specific fish predation on the survival of invasive and native gammarids. J Sea Res. 2010; 64: 369–372. https://doi.org/10.1016/j.seares.2010.05.006

47. Deagle BE, Tollit DJ. Quantitative analysis of prey DNA in pinniped faeces: Potential to estimate diet composition? Conserv Genet. 2007; 8: 743–747. https://doi.org/10.1007/s10592-006-9197-7

48. Polz MF, Cavanaugh CM. Bias in template-to-product ratios in multitemplate PCR. Appl Environ Microbiol. 1998; 64: 3724–3730. PMID: 9758791

49. Piñol J, Mir G, Gomez-Polo P, Agustí N. Universal and blocking primer mismatches limit the use of high-throughput DNA sequencing for the quantitative metabarcoding of arthropods. Mol Ecol Resour. 2015; 15: 819–830. https://doi.org/10.1111/1755-0998.12355 PMID: 25454249

50. Vestheim H, Deagle BE, Jarman SN. Application of blocking oligonucleotides to improve signal-to-noise ratio in a PCR. Methods Mol Biol. 2011; 687: 265–74. https://doi.org/10.1007/978-1-60761-944-4_19 PMID: 20967615

51. Vestheim H, Jarman SN. Blocking primers to enhance PCR amplification of rare sequences in mixed samples–a case study on prey DNA in Antarctic krill stomachs. Front Zool. 2008; 5: 12. https://doi.org/10.1186/1742-9994-5-12 PMID: 18638418

52. Oehm J, Thalinger B, Mayr H, Traugott M. Maximizing dietary information retrievable from carcasses of Great Cormorants *Phalacrocorax carbo* using a combined morphological and molecular analytical approach. Ibis (Lond 1859). 2016; 51–60. https://doi.org/10.1111/ibi.12337 PMID: 26877544

53. Bowser AK, Diamond AW, Addison JA. From Puffins to Plankton: A DNA-Based Analysis of a Seabird Food Chain in the Northern Gulf of Maine. Stow A, editor. PLoS One. 2013; 8: e83152. https://doi.org/10.1371/journal.pone.0083152 PMID: 24358258

54. Sheppard SK, Bell J, Sunderland KD, Fenlon J, Skervin D, Symondson WOC. Detection of secondary predation by PCR analyses of the gut contents of invertebrate generalist predators. Mol Ecol. 2005; 14: 4461–4468. https://doi.org/10.1111/j.1365-294X.2005.02742.x PMID: 16313606

55. Hunter JR, Kimbrell CA. Egg cannibalism in the northern anchovy, *Engraulis mordax*. Fish Bull US. 1980; 78: 811–816.

56. Meier HEM. Baltic Sea climate in the late twenty-first century: a dynamical downscaling approach using two global models and two emission scenarios. Clim Dyn. 2006; 27: 39–68. https://doi.org/10.1007/s00382-006-0124-x

57. Lefébure R, Larsson S, Byström P. Temperature and size-dependent attack rates of the three-spined stickleback (*Gasterosteus aculeatus*); are sticklebacks in the Baltic Sea resource-limited? J Exp Mar Bio Ecol. Elsevier B.V.; 2014; 451: 82–90. https://doi.org/10.1016/j.jembe.2013.11.008

Chapter 8

The pitfalls of biodiversity proxies: Differences in richness patterns of birds, trees and understudied diversity across Amazonia

Ritter et al. 2019, Scientific Reports

www.nature.com/scientificreports

SCIENTIFIC REPORTS
nature research

OPEN

The pitfalls of biodiversity proxies: Differences in richness patterns of birds, trees and understudied diversity across Amazonia

Camila D. Ritter[1,2,3*], Søren Faurby[2,3], Dominic J. Bennett[2,3], Luciano N. Naka[4], Hans ter Steege[5,6], Alexander Zizka[7], Quiterie Haenel[8], R. Henrik Nilsson[2,3,10] & Alexandre Antonelli[2,3,9,10]

Most knowledge on biodiversity derives from the study of charismatic macro-organisms, such as birds and trees. However, the diversity of micro-organisms constitutes the majority of all life forms on Earth. Here, we ask if the patterns of richness inferred for macro-organisms are similar for micro-organisms. For this, we barcoded samples of soil, litter and insects from four localities on a west-to-east transect across Amazonia. We quantified richness as Operational Taxonomic Units (OTUs) in those samples using three molecular markers. We then compared OTU richness with species richness of two relatively well-studied organism groups in Amazonia: trees and birds. We find that OTU richness shows a declining west-to-east diversity gradient that is in agreement with the species richness patterns documented here and previously for birds and trees. These results suggest that most taxonomic groups respond to the same overall diversity gradients at large spatial scales. However, our results show a different pattern of richness in relation to habitat types, suggesting that the idiosyncrasies of each taxonomic group and peculiarities of the local environment frequently override large-scale diversity gradients. Our findings caution against using the diversity distribution of one taxonomic group as an indication of patterns of richness across all groups.

Despite significant advances in our understanding of global biodiversity, a fundamental question remains poorly understood[1]: *Do the same ecological patterns apply to macro and micro-organisms?* In fact, our understanding of biodiversity is biased towards charismatic and relatively easily identifiable taxa. For instance, for birds and flowering plants, an estimated 98%[2,3] and 69%[3] respectively of the extant species have been formally described. Yet, even in these taxonomically well-described groups, the geographic distribution of many species remains poorly understood (the 'Wallacean shortfall'[4]). The overwhelming majority of the extant biodiversity, however, does not belong to these groups. All vertebrates combined represent only 0.7%, and all flowering plants only 3%, of the total estimated number of eukaryotic species. Many species, particularly of invertebrates and micro-organisms, are yet to be described (the 'Linnaean shortfall'[4]) and their distribution has yet to be documented.

A pre-requisite to overcoming these shortfalls is the ability to record and recognize species. Species identification, however, requires taxonomic expertise, which in turn requires a substantial and long-term investment of resources, time and infrastructure, especially when species are vouchered and deposited in natural history collections[5]. Recently, high-throughput DNA sequencing approaches, in combination with DNA metabarcoding[6], have

[1]Department of Eukaryotic Microbiology, University of Duisburg-Essen, Universitätsstrasse 5 S05 R04 H83, D-45141, Essen, Germany. [2]Gothenburg Global Biodiversity Centre, Box 461, SE-405 30, Göteborg, Sweden. [3]Department of Biological and Environmental Sciences, University of Gothenburg, Box 463, SE-405 30, Göteborg, Sweden. [4]Laboratório de Ornitologia, Departamento de Zoologia, Universidade Federal de Pernambuco, Recife, PE, Brazil. [5]Naturalis Biodiversity Center, Leiden, Netherlands. [6]Systems Ecology, Free University, Amsterdam, Netherlands. [7]German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Deutscher Platz 5e, 04103, Leipzig, Germany. [8]Zoological Institute, University of Basel, Vesalgasse 1, CH-4051, Basel, Switzerland. [9]Royal Botanic Gardens, Kew, TW9 3AE, Richmond, Surrey, UK. [10]These authors contributed equally: R. Henrik Nilsson and Alexandre Antonelli. *email: kmicaduarte@gmail.com

enabled the identification of organisms and the estimation of diversity from bulk (unsorted) biological samples, facilitating the documentation of spatial diversity patterns across the tree of life[7,8].

Besides geographic differences, large-scale biodiversity patterns vary among taxonomic groups. Some studies have already assessed the correlations between the diversity of macro and micro-organisms. On a global scale, a mismatch of diversity was found between below-ground organisms (bacteria, fungi and mesofauna) and above-ground organisms (mammal, birds, amphibians and vascular plants)[9]. Furthermore, bacterial diversity was higher in temperate regions, while fungi showed a weak latitudinal pattern[10]. In another study, fungal diversity displayed a latitudinal gradient but was uncorrelated with plant diversity[11]. For Neotropical forests, protists showed the same pattern of diversity as macro-organisms[12], and fungi and bacteria followed the elevational gradient of diversity in the Andes[13]. The pattern of richness of fungi and bacteria in the mineral soil, however, was different from that of plants, not linear, with fungi having the lowest richness in median elevation and bacteria the highest[13]. In this context, the congruence or divergence in diversity across taxa remains unclear. This is problematic, since micro-organims are the most diverse and abundant groups in any habitat[14] and are essential for ecosystem function[15] and the fitness of higher organisms[16], meaning that general insights into the distribution and drivers of diversity require their inclusion[17].

Although insufficient biological knowledge prevails in nearly all ecosystems around the world, this problem is most conspicuous in tropical environments, and in particular in tropical forests. Amazonia is the world's largest and most biodiverse tropical forest. On a large spatial scale, most macroscopic taxa show consistent patterns of diversity, possibly as a response to abiotic conditions and processes[18–20]. In this region, one of the most conspicuous patterns of species richness in well-studied groups, such as birds and trees, is a west-to-east diversity gradient: from the highly diverse areas on the eastern Andean slopes to the relatively less diverse areas on the Guiana Shield in the north and eastern Amazonian lowlands[18–22]. Several explanations for this pattern have been suggested, including the effects of marine incursions[23–25], bedrock geology[26], mountain base formation[18], soil fertility[18,27] and, more recently, a diversification process driven by moisture[28].

While most of Amazonia is covered by lowland non-flooded terra-firme forests, several other vegetation types, such as flooded forests or white sand ecosystems, are common and widespread throughout the basin. Patterns of plant and avian diversity vary dramatically with vegetation type; as a general rule, terra-firme forests are more diverse than seasonally flooded forests[29–31]. Forests that are seasonally flooded by nutrient-rich, white-water rivers (várzeas) are more diverse than forests seasonally flooded by acidic, nutrient-poor black-water rivers (igapós[31,32]). Finally, both types of flooded forests are more diverse than naturally open areas on nutrient-impoverished sandy soils (campinas[31,33–36]). The drivers of these patterns remain elusive but may be associated with geological processes, soil fertility, inundation gradient, type of water[37] and also with the size and fragmented distribution of these "smaller vegetation types" on which the colonization of species may be in part attributed to chance[38,39]. However, such patterns could in principle be specific to plants and vertebrates. Other taxa, such as fungi, bacteria and other micro-organisms could display different diversity patterns. Indeed, in a previous study using part of our data, we found different patterns for micro-organismal richness among Amazonian habitat types[40], but a similar pattern of higher terra-firme diversity than campina diversity was found for fungi in Colombian Amazonia[41]. The contrasting patterns between micro- and macro-organisms may have major implications for our understanding of general diversity patterns and for conservation.

In this study, we test whether patterns of tree and avian species richness are similar to those found in Operational Taxonomic Units (OTU[42]) mainly targeting micro-organisms. For this purpose, we compare OTU richness generated from environmental sequencing in four Amazonian localities, with richness estimates from existing taxonomic inventories for trees and birds in the vicinity (Fig. 1). For the OTU analyses, we examine three different sample types (soil, litter and insect bulk samples) and three sequence markers (the ribosomal 16S, 18S and the mitochondrial COI, which target prokaryotes, eukaryotes and metazoans, respectively). We test if large-scale diversity patterns known from plants and birds (increasing richness from east-to-west and from campinas to flooded forests and to terra-firme forests) can be recovered with our OTU and inventory data. If OTUs and traditional taxonomic species richness show approximately the same diversity patterns, metabarcoding could offer a rapid and cost-effective alternative for biodiversity assessments, without the demand for taxonomic expertise. In that case, the detection and protection of high diversity areas would be facilitated[43–45], and taxonomists could focus on species descriptions and other important directions of research, rather than spending time on routine identifications. If, however, OTU richness and species richness are decoupled, the idiosyncrasies of each taxonomic group would make generalizations difficult and call into question our current understanding of the distribution of biodiversity in the world's largest rainforest. Importantly, a rapid and reliable assessment of Amazonian diversity is increasingly crucial, as deforestation rates are currently escalating to alarmingly high levels[46].

Results

After rarefaction, we obtained a total of 15,563 OTUs for 16S; 17,017 for 18S; and 14,964 for COI (see Supplementary Table S1 for the DNA concentration, number of reads, number of OTUs and Shannon estimate for each plot). The taxa with the highest number of identified OTUs across all samples were: Alphaproteobacteria (15%), Acidobacteria (10%), Planctomycetes (10%), Bacteroidetes (10%), Actinobacteria (10%) and Chloroflexi (10%) (Fig. 2A) for prokaryotes (the 16S marker); and Fungi (20%, mainly Ascomycota and Basidiomycota), Cercozoa (15%) and Alveolata (10%) (Fig. 2B) for eukaryotes (the 18S marker). For the COI marker, the taxa with the highest number of OTUs were Fungi (30%, mainly Ascomycota and Basidiomycota) followed by Hexapoda (10%; Fig. 2C). The proportion of unclassified OTUs was around 10%, 25% and 40% for 16S, 18S and COI, respectively, reflecting the incompleteness of public databases for these markers, beyond the possible sequence errors/chimeras. The lack of representative sequences is more problematic for COI, since usually this marker is sequenced just for metazoans.

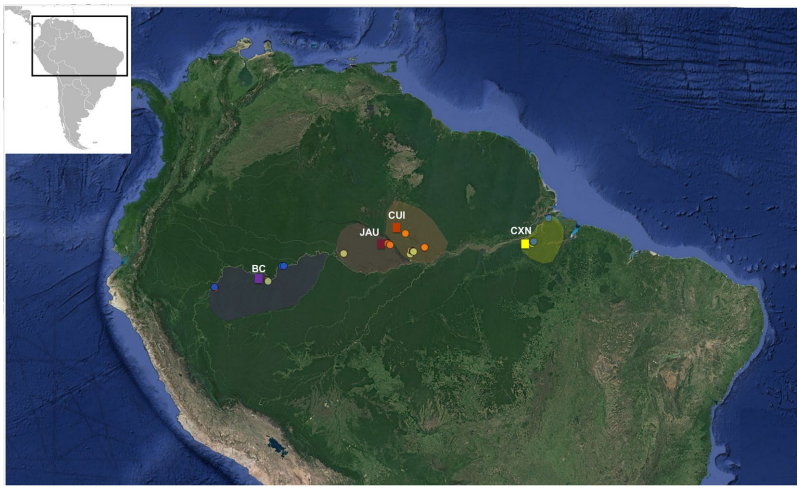


Figure 1. Map of sampling localities. Circles represent plots pertaining to the Amazon Tree Diversity Network (ATDN) used in this study, which represent different forest types: igapós (orange), várzeas (blue) and terra-firme (green). The semi-transparent polygons show the interfluves from which those plots were selected. Squares represent the locations of the metabarcoding sampling that were compared to the ATDN data. In each locality, we sampled different habitats: in Benjamin Constant (BC) we sampled terra-firme, igapós and várzeas; in Jaú (JAU) and Cuieras (CUI) we sampled terra-firme, campinas and igapós. At each of the three localities we sampled nine plots. In Caxiuanã we sampled terra-firme, campinas, várzeas and igapós, totaling 12 plots. The map was contructed with Qgis v.3.6.2[96].

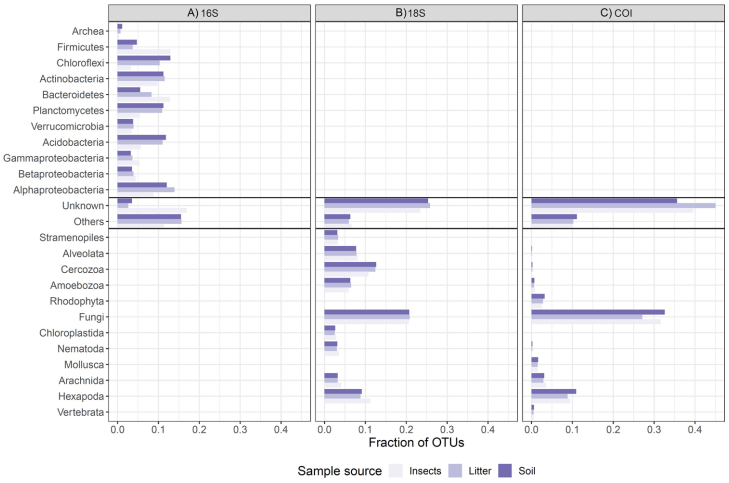


Figure 2. Taxonomic composition of OTU communities. The plots show the breakdown of OTUs into taxonomic groups from (**A**) 16S, (**B**) 18S and (**C**) COI, respectively, coloured by sample type. There is no clear taxonomic variation between soil and litter samples other than some variation in the taxonomic composition for insect samples in the 16S and 18S data sets.

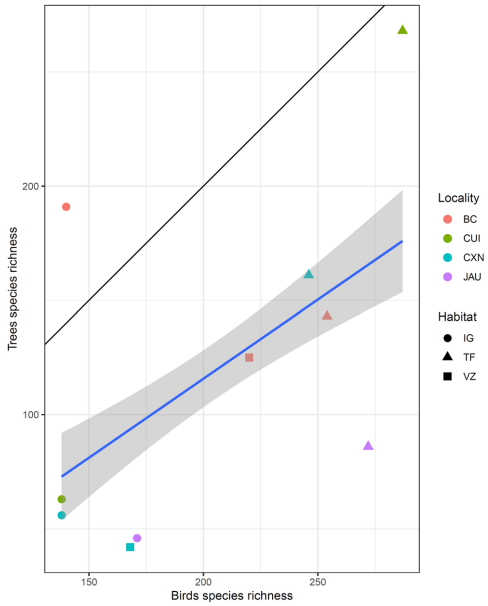


Figure 3. Regression between plot-level species richness of trees and regional species richness of birds for the localities sampled. The thick blue line shows the linear regression with standard error indicated by the shaded area. The thin solid black line shows x = y (perfect correlation). There is a weak but significant relationship between the species richness of these two taxonomic groups (posterior mean = 0.01; $p < 0.001$). The colour represents the localities: BC = Bejamin Constant, CUI = Cuieras, CXN = Caxiuanã, JAU = Jaú. The symbols represent the habitat type: IG = igapós. TF = terra-firmes, VZ = várzeas.

Regional species richness for birds was poorly related to the average plot-level species richness for trees (posterior mean = 0.01, $p < 0.001$; Fig. 3). When divided by habitat, the regressions were significant for terra-firme ($adjR^2 = 0.21$, $p = 0.002$) and igapó ($adjR^2 = 0.11$, $p = 0.03$). Only two datapoints were available for várzeas.

The average species richness of trees (1 ha plot-level), plot-level DNA-based OTU richness and regional bird species richness all decline along a west-to-east gradient (Table 1; Fig. 4). Species and OTU richness were generally decoupled across vegetation types. The species richness of birds and trees showed the richness gradient terra-firme > várzea > igapó > campinas did not show the same gradient among vegetation types, with campinas having the highest richness (Table 1; Fig. 4). The number of species (trees and birds) and DNA-based OTUs per habitat in each locality is available in Supplementary Table S2.

The relationship between plot-level DNA-based OTU and plot-level tree and regional bird richness was not significant (posterior mean = 0, $p > 0.05$ for both tests that were analyzed separately; Table 2). Only the metabarcoding predictors (sample and marker type) were significant in both models (plot-level DNA-based OTU richness *versus* nearby plot-level tree richness and plot-level DNA-based OTU richness *versus* regional birds richness; Table 2). We found the same pattern when we subdivided the metabarcoding data based on taxonomy (prokaryotes, protists, fungi and metazoan; Table S3). The random effects of "locality" and "habitat" type were not significant. We found a significant positive relationship between plot-level DNA-based OTU richness and plot-level species richness of nearby tree plots (eight positive regressions out of nine tests; $p = 0.039$; Table 3; Fig. 5) when considering a binomial distribution. In contrast, there was no clear relationship between plot-level OTU richness and regional bird species richness (five negative regressions out of nine tests; two-tailed probability 0.51; Table 3; Fig. 5).

Discussion

Our study indicates that OTU and species richness shows a declining west-to-east diversity gradient, yet the biodiversity patterns of macro- and DNA-based OTUs are largely decoupled across Amazonia. We found no relationship between DNA-based OTU richness estimated from metabarcoding of environmental samples and species richness estimated from previous field inventories. These results suggest that at the regional scale, the diversity distribution of one taxonomic group should not be used as a general proxy for diversity of another, nor as an indication of overall patterns of richness. At small spatial scales, the idiosyncrasies of each taxonomic group and the peculiarities of each environment appear to be more important than general diversity patterns, which differ among organism types.

	Meta (OTUs)	16S (OTUs)	Protists 18S (OTUs)	Protists COI (OTUs)	Fungi 18S (OTUs)	Fungi COI (OTUs)	Metazoa 18S (OTUs)	Metazoa COI (OTUs)	Birds (regional species)	Trees (average species, 1 ha plot)
Locality										
Benjamin Constant	**907**	**1525**	**262**	18	**263**	83	**205**	38	**205**	152
Jaú	813	1336	213	32	212	125	163	65	203	86
Cuieras	714	1074	199	28	200	126	155	61	194	66
Caxiuanã	877	1338	220	**39**	222	**157**	171	**84**	170	**166**
Habitat										
Terra-firme	808	1266	214	32	215	127	166	63	**265**	**164**
Várzea	843	1358	234	21	236	106	**187**	56	194	82
Igapó	757	1212	215	19	214	97	170	44	147	89
Campina	**973**	**1511**	**239**	**48**	**241**	**176**	178	**95**	150	N/A

Table 1. Mean number of all OTUs ('meta'; comprising prokaryotes and eukaryotes), OTUs by taxonomic groups and species ('birds' and 'trees') for locality and habitat. OTUs were divided by the main taxonomic group; 16S comprises mostly bacteria, and 18S and COI were divided into protists, fungi and metazoan. For localities, the mesuared richness shows a gradient from west to east: Benjamin Constant > Jaú > Cuieras > Caxiuanã. For habitat type, the measured richness reflects the order expected based on the literature for macro-organisms: terra firme > várzea > igapó > campinas. The highest richness in each category is highlighted in bold. The patterns are different from our expectations for localities, with Caxiuanã being richer than expected for metabarcoding and trees. For habitats, OTU richness is also different from the expected, but for birds and trees the species richness reflects the previously documented pattern. Tree richness is not reported for campinas since it does not capture the known flora of those habitats and is dominated by other growth forms (e.g., herbs and shrubs).

It is important to acknowledge that we compared data aggregated at different spatial scales and generated using different methods in order to produce the richness estimates used here. In addition, there are differences in the exact locations of the trees surveyed and the metabarcoding plots sampled for this study. These considerations make a direct comparison of richness challenging and worth further exploration by future studies based on primary inventories. However, our primary aim was to assess correlations between *proxies* of species richness. This means that despite these challenges, if the regional-scale processes are important (locality, habitat type), the levels of alpha diversity should increase as a function of the source pool (unsaturated type I relation[47,48]). Therefore, if the west-to-east gradient or habitat differences hold true for all samples, a positive and significant relationship should be found across our data sets. If not, this would suggest that other factors may be more important in determining richness from local to regional scales.

For prokaryotes, diversity is often high in pastures and agricultural fields, which generally have low animal and plant diversity at the local to regional scale[49–53]. However, some bacterial groups, such as the Alphaproteobacteria and Planctomycetes[53], are more diverse in undisturbed forests. Both of these groups were abundant in our samples, accounting for 35% of our 16S data (Fig. 2A). As a result, when looking for general patterns of richness in bacteria, a negative correlation with trees and birds could be expected, but these effects could be masked by other groups that are positively correlated with macro-organisms, as is the case in Alphaproteobacteria and Planctomycetes.

Patterns of diversity can be distinct for different taxonomic groups, and the wide taxonomic range of metabarcoding studies can mask taxon-specific patterns. Furthermore, different markers target different species and may have added some noise in our analysis. For instance, for fungi in litter samples, 18S and COI displayed the opposite pattern (Fig. S1). Previous studies have reported a decoupling between fungi and plant diversity worldwide[11] whereas others have found a positive relationship[13,41] and a similar community turnover[54]. For other groups, such as insects, diversity is often positively correlated with plant diversity[55,56]. Additionally, soil protists can have similar biogeographic patterns to macro-organisms in lowland Neotropical rainforests[12], which is expected to have a positive effect in the regression of protist OTUs and tree and bird species richness. Our data showed similar patterns overall for metazoans, fungi and protists for these same markers (Table S2, Fig. S1). However, our results highlight the need for further exploration of biotic interactions and diversity metrics, as contrasting results can be found within the same taxonomic groups (e.g. fungi sequenced with 18S and COI, Fig. S1).

A west-to-east decline in diversity has repeatedly been documented in birds[57,58] and plants[20,21,57] for Amazonia and is also partly reflected in our metabarcoding data, other than for the easternmost locality (Table 1). A positive correlation with this diversity should be found if all groups shared the same overall diversity pattern due only to the same abiotic conditions (e.g. moisture[28], nutrient levels[59] or geology[26]), yet regional and local deviations appear idiosyncratic among taxa. For instance, the combined data from the Amazon Tree Diversity Network across the entire Amazon basin cleary show a west-to-east diversity gradient, but contain multiple outliers in the eastern part of the Negro River close to the Cuieiras area surveyed here[21]. This is consistent with the observed higher-than-expected tree richness in terra-firme from Cuieras as revealed from our data (Fig. 4D) and this may have affected the results of our regressions due to our limited sampling. In addition, Benjamin Constant has the poorest bird inventories, possibly resulting in underestimated richness for this area in our data.

By adding more data and analyses to our previous study[40], we could provide further evidence that the plot-level DNA-based OTU richness gradient differs from the plot-level tree species richness and from the regional bird species richness across vegetation types. The general richness pattern for vertebrates and plants, also reported here with our bird and tree data, is: terra-firme > várzea > igapó > campina[21,30–36]. However, we found that campinas

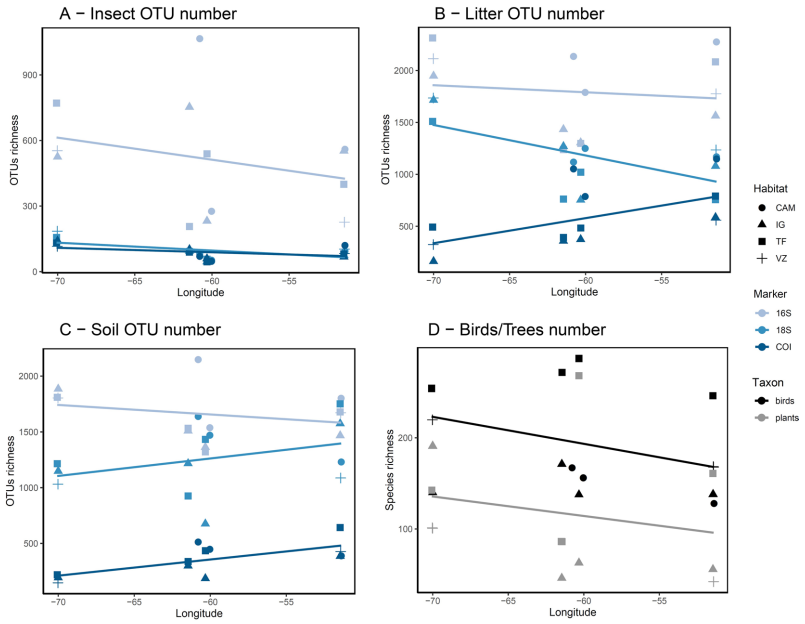


Figure 4. Metabarcoding OTU and species richness of birds and trees per longitude and habitat type. The plots show OTU richness measured from metabarcoding samples of (**A**) insects, (**B**) litter and (**C**) soil. Plot (**D**) shows the known species richness for trees and birds from which those samples were obtained. The colour-coding in A–C indicates marker type and in D the taxonomic group and the symbols indicate habitat type (CAM: campinas, IG: igapó, TF: terra firme and VZ: várzea). The results for A–C indicate that OTU richness varies significantly with location and habitat type, with the highest overall richness obtained from 16S data. For species richness of trees and birds, a consistency between environment richness (TF > VZ > IG > CAM) can be observed, and a west-to-east gradient, as generally expected based on large-scale inventories. For OTUs, an overall pattern with the highest richness in campinas is observed. The west-to-east gradient is observed in general, except for COI litter and 18S and COI soil.

make up the richest habitat in our OTUs data (Table 1, Fig. 4). This habitat is usually considered less diverse for macro-organisms than more forested habitats in Amazonia, such as terra-firme and flooded forests[18,20,21,33,36], a relationship confirmed for Colombian Amazonian fungi[41]. Previous studies have reported on the importance of campinas for beta-diversity[36], but these habitats have long been considered species-poor environments[60]. In contrast, our results suggest that these environments may be hyperdiverse for microbial diversity (Fig. 4A–C). However, we note that campinas have an insular distribution in Amazonia, being surrounded by a "sea" of terra-firme forests[61,62]. OTU diversity in these patches could, potentially, be over-estimated due to DNA transported from nearby forest species, for instance through leaves, fungal spores and other debris[63]. This effect will be hard to test, but it is important to stress that the community composition of campinas was significantly different from the other habitats[40,64] and there is a rich micro-organismal community that is genuinely from campinas.

The different spatial scales for our analyses – plots of 28 m of radius for metabarcoding data, 1 ha plots for trees and species pools in the interfluvia for birds, influences our species richness comparison. However, within each taxonomic group, the species richness should be consistent across these scales if the west-to-east and habitat gradients are the dominating factors explaining the richness gradient. The outliers in our data (e.g. tree richness in Cuieras and OTU richness in campinas) may have had the strongest effect in the general regression between the OTUs and species richness for birds and for trees. For trees, the pattern we recovered reflects outliers already identified in a previous study[21]. These considerations suggest that even with the different spatial scales used here and in other studies, if the west-to-east gradient was the strongest factor explaining diversity, it should produce a positive correlation. However, the outliers showed that the specifity of localities affected the general pattern.

There are still numerous uncertainties in the underlying biodiversity data and in our ability to generalize overall diversity patterns and identify their main determinants from local to regional scales. We therefore recommend the further validation of the patterns reported here through the generation and analysis of independent data, sampled under standardised conditions for multiple organism groups. With a standardized protocol and additional analyses, such as, for example, that of the metatranscriptome[65] to target only metabolically active organisms, it

Taxon	Effect	Variables	post.mean	l-95% CI	u-95% CI	eff.samp	pMCMC
Trees	Fixed	Richness taxa	0.00	0.00	0.00	693.3	0.464
		Marker 16S	**5.73**	5.47	6.04	1000.00	<0.001
		Marker 18S	**5.00**	4.67	5.28	1000.00	<0.001
		Marker COI	**4.17**	3.88	4.47	1000.00	<0.001
		Sample Litter	**1.80**	1.58	2.02	1136.00	<0.001
		Sample Soil	**1.68**	1.45	1.91	1000.00	<0.001
	Random	Locality	0.14	0.00	0.07	107.5	NA
		Habitat	0.00	0.00	0.00	1000.00	NA
Birds	Fixed	Richness taxa	0.00	0.00	0.00	873.78	0.902
		Marker 16S	**5.81**	5.39	6.20	1000.00	<0.001
		Marker 18S	**5.03**	4.66	5.44	1000.00	<0.001
		Marker COI	**4.29**	3.90	4.69	1000.00	<0.001
		Sample Litter	**1.89**	1.67	2.10	1000.00	<0.001
		Sample Soil	**1.75**	1.53	1.95	1000.00	<0.001
	Random	Locality	0.01	0.00	0.04	338.1	NA
		Habitat	0.01	0.00	0.02	711.6	NA

Table 2. Coefficients for the general linear model fitted in a Bayesian framework using Markov chain Monte Carlo (MCMC) methods for OTU richness against species richness of trees and birds. The model was adjusted with the Poisson family distribution considering taxonomic richness, marker and sample type as fixed effects, and locality and habitat type as random effects. For trees and birds, the taxonomic richness is not significant, whereas the marker and sample type are. Significant values of the post mean of the coefficients (at $p < 0.05$) are shown in bold.

Taxon	Sample type	16S	18S	COI
Trees	Insect	**0.26**	**0.01**	**0.004**
	Litter	**0.17**	**0.13**	*−0.02*
	Soil	**0.09**	**0.38**	**0.18**
Birds	Insect	**0.09**	*−0.08*	*−0.10*
	Litter	*−0.08*	*−0.18*	**0.29**
	Soil	*−0.01*	0.23	**0.32**

Table 3. Results for the generalized linear mixed effects models considering each marker and sample type separately. For each model, the coefficient is presented. No single regression is significant after Bonferroni correction for multiple tests ($p < 0.00275$). In order to illustrate the pattern in the sign of the effect, we have given all positive slopes in **bold** and all negative ones in *italics*. It is evident that the vast majority of slopes are positive between OTU and tree richness (8/9 $P = 0.039$), while there is no consistency for the relationship between OTU and bird richness (4/5 n.s.).

will be possible to avert these shortcomings and to draw stronger conclusions on species interactions[66,67], abiotic diversity drivers[64,68] and above-ground and below-ground feedback[69].

A recent global study comparing below-ground organisms (bacteria, fungi and mesofauna) with above-ground organisms (mammal, birds, amphibians and vascular plants) found a diversity mismatch of 27%[9]. The findings from this and previous studies that micro- and macro-organismal diversity are often decoupled has important implications for conservation. It is genuinely worrying in the context of bioversity loss[70], since a large proportion of the world's biodiversity may be lost without notice, particularly in Amazonia[46]. Micro-organisms are essential for ecosystem functioning, as they constitute the majority of the diversity of any ecosystem. As highlighted by O'Malley & Dupré[17], the excessive focus on macro-organisms may have distorted our understanding of general patterns of biodiversity. There is therefore a danger that conservation strategies may be inadequate, if their primary focus is to maintain ecosystem functionality and the biotic interactions[71].

Conclusions

In this study, we found that other than displaying a declining west-to-east gradient at large spatial scale, species richness patterns are not consistent across taxa in Amazonia. In particular, patterns in the diversity of micro-organisms (which comprise the bulk of the total diversity) differ strongly from patterns in birds and plants, particulary in connection with habitat type. Furthermore, we found large differences in species richness and diversity patterns between i) metabarcoding of environmental samples and nearby taxonomic inventories, and ii) different genetic markers used for DNA barcoding. Importantly, our results suggest that diversity patterns differ considerably among taxonomic groups, making the use of single taxa as a proxy for total diversity problematic, especially for conservation purposes. This study highlights the importance of integrative and data-rich approaches to studying and describing diversity.

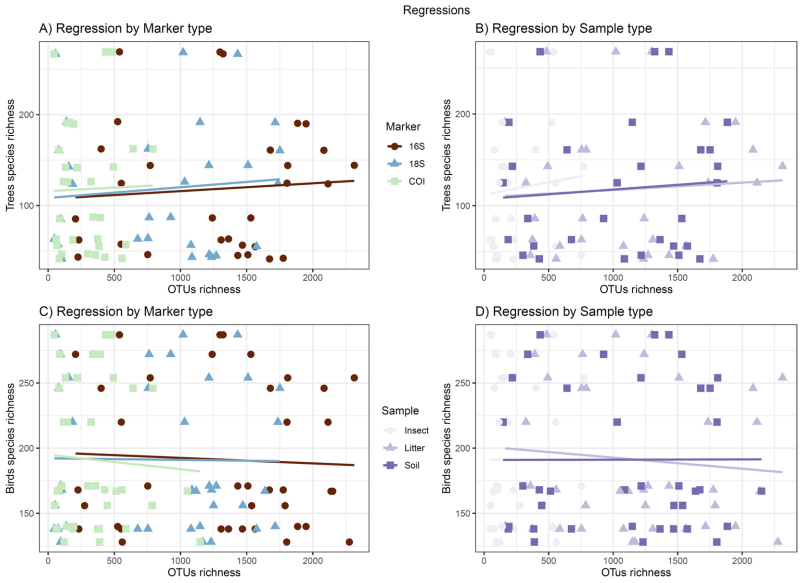


Figure 5. Regression between OTU and species richness. The lines show the regressions between OTUs and tree richness in (**A**, **B**) and between OTU and bird richness in (**C**, **D**). The samples are coloured per marker in (**A, C**) and per sample type in (**B, D**). The vast majority of slopes are positive between OTU and tree richness. However, for birds there is no consistency in the relationship between DNA-based OTU and bird species richness.

Material and Methods

Study areas. We sampled four localities across a west-to-east transect in Brazilian Amazonia (Fig. 1[40]). These areas include: Benjamin Constant (a municipality which is the westernmost locality in our sampling scheme, located south of the Amazon river); Jaú (a national park in central Amazonia situated west of the Negro river and north of the Amazon river); Cuieras (a biological reserve east of the Negro river and north of the Amazon river); and Caxiuanã (the easternmost locality in our sampling: a national forest situated south of the Amazon river; Fig. 1). We chose these localities to maximize geographic distance and to cover all major vegetation types, i.e. terra-firme, várzeas, igapós and campinas (see ref. [40] for a more detailed description of the locations surveyed).

Sampling of metabarcoding data. We collected mineral soil, litter (the organic matter above the mineral soil) and insects in three plots in each major vegetation type present at each locality (3 to 4 depending on the locality; see ref. [40] for more details) in November 2015. First, we installed a SLAM trap in the middle of each plot. SLAM traps are dome-shaped, tent-like insect traps made of fine mesh-netting, widely used in entomological studies and aimed at capturing strong-flying insects that typically fly upwards after hitting a fine-scale net (e.g. wasps, mosquitos and butterflies). These insects were ultimately trapped in a bottle filled with ethanol at 96% concentration. The traps were kept open for 24 hours in each plot. After capture, the insects were preserved in a clean plastic bottle with new 96% ethanol until DNA extraction.

We sampled soils and litter following Tedersoo *et al.*[11] to minimize information loss while keeping comparability between this and other large-scale studies. First, 20 trees were randomly selected within a 28 m radius of each SLAM trap. To reduce the risk of contamination, we wore gloves and a nose-and-mouth mask and replaced the gloves between each sampled tree. We sampled litter and soil cores in opposite directions of each selected tree. In total, 40 soil and 40 litter samples were collected per plot. The soil and litter samples were subsequently pooled into one combined soil and one combined litter sample for each plot. The litter consisted of all organic material above the mineral soil and varied from 0–50 cm in thickness. We then collected soil in the same places, with the samples taken from the top 5 cm of the mineral soil using a metal probe with a 2.5 cm diameter. The soil probe was sterilized with fire after collecting soil from both sides of each tree to prevent cross-contamination between samples. The samples were stored in plastic bags with the same weight of sterilized white silica gel (14 mm silica grain size). The silica was pre-treated for two minutes in a microwave oven (800 W) and exposed to 15 min of UV light to prevent contamination in our samples from any micro-organisms present in the silica. All plots were tagged with GPS coordinates. All dry soil, litter samples and ethanol insect samples were processed at the University of Gothenburg, Sweden. For more details of the collection protocol, see ref. [43].

DNA extraction. For soil, 10 g (dry weight) of each sample and 15 ml of each litter sample (corresponding to 3–10 g of dry weight litter, depending on texture and composition of each sample) and a negative control were processed for total DNA extraction using the PowerMax® Soil DNA Isolation Kit (MO BIO Laboratories), according to the manufacturer's instructions (see details in ref. [40]). For insects, we followed the non-destructive protocol described in Aljanabi and Martinez[72], we also included a negative control for insect extractions.

PCR Amplification. We used three genetic markers to target different organisms: 16S for prokaryotes, 18S and COI for eukaryotes in general. For amplification of ribosomal small subunit (SSU) 18S rRNA in soil and litter samples, we targeted the V7 region of the gene using the forward and reverse primers (5′-TTTGTCTGSTTA ATTSCG-3′) and (5′-TCACAGACCTGTTATTGC-3′) designed by Guardiola *et al.*[73] to yield 100 to 110 base pair (bp) fragments (see details in ref. [27]). For the ribosomal small subunit (SSU) 16S rRNA, we targeted the V3–V4 region (~460 bases) of the 16S rRNA gene using the forward primer (5′-CCTACGGGNGGCWGCAG-3′) and reverse primer (5′-GACTACHVGGGTATCTAATCC-3′) from Klindworth *et al.*[74]. For the cytochrome c oxidase subunit I mitochondrial gene (COI), we amplified a region of ~313 bases using an internal forward primer (5′-GGWACWGGWTGAACWGTWTAYCCYCC-3′[75]) and the COI degenerate reverse primer (5′-TAAACTTCA GGGTGACCAAARAAYCA-3′[76]). Amplification and sequencing were carried out by Macrogen (Republic of Korea) following standard protocols using the Illumina MiSeq. 2 × 250 (18S) and 2 × 300 (16S and COI) platform, including the negative control to check possible sequences errors and cross-sample contaminations[77]. Part of the data presented here has already been published. The soil and litter data for 16S and 18S were already analysed in previous studies[40,64]. Soil for COI and insect samples for the three markers were previously analysed in Benjamin Constant[43]. Here we present new data for COI for litter (all data), and COI for soil; as well as16S, 18S and COI for insects for Jaú, Cuieras and Caxiuanã. All raw sequences are available in GenBank under Bioproject PRJNA464362.

Sequence analyses and taxonomic assessments. We used the USEARCH/UPARSE v9.0.2132 Illumina paired reads pipeline[78] to merge the paired sequences with a maximum of five mismatches allowed, truncate by the length (80 bp for 18S, 400 bp for 16S and 290 bp for COI), filter sequence reads for quality and discard reads with >1 total expected errors for all bases in the read after truncation, de-replicate and sort reads by abundance, infer OTUs by 97% of similarity and remove singletons. We filtered the data to discard artificial sequences (e.g. chimeras), and we clustered sequences into OTUs at a minimum similarity of 97% using a "greedy" algorithm that performs chimera filtering and OTU clustering simultaneously[78]. We address all OTUs registered in the negative controls (18S = 595 OTUs, 16S = 379 OTUS, negative control fail in sequencing for COI) and excluded them from our data sets (Tables S4. and S5). For 16S and 18S data, we used SILVA 1.3[79] for assessment of the taxonomic composition of the OTUs, using a representative sequence from each OTU as query sequence and the SINA v1.2.10 reference data for ARB SVN (revision 21008[80]) for local BLAST searches[81] of both markers. As reference COI data, we used all COI sequences deposited in GenBank until August 2018[82] in our BLAST searches. All searches were conducted with the same criterion: a minimum 80% similarity and an e-value of 0.001.

Compilation of taxonomic data. We compared the OTU diversity estimated from our environmental samples with morphology-based taxonomic estimates of species richness for trees and birds. For trees, we used the data from the Amazon Tree Diversity Network (http://atdn.myspecies.info/). That project links plots across all Amazonia from different vegetation types, where a full inventory was made of all free-standing trees up to 10 cm in diameter at breast height (dbh). Trees were identified to the level of species or morphospecies. We compiled the mean richness of tree species in all 1-ha plots within each ecosystem type and interfluvial for which we had metabarcoding data (Fig. 1). For two plots that had an area of 1.3 ha, we estimated the number of individuals expected in 1 ha (number of individuals / 1.3). We then rarefied the plot by the number of expected species using the "rarefy" function in the package vegan v. 2.4–3[83] in R v3.3.2[84]. Since trees are only a minor component of the vegetation in campinas[85], we considered them a poor proxy for plant diversity in such plots. We therefore excluded campinas in the analyses of the relationship between trees and OTU richness.

For birds, we used published compilations for our study sites whenever available. This was the case for Jaú National Park[33,86], and Caxiuanã National Forest[87]. For Cuieiras, we used data from Manaus, a well-studied nearby locality[88] that is situated in the same interfluvium area and should therefore have a very similar species pool. For Benjamin Constant, which lacks available published sources, we created a hypothetical species list based on data from the Global Biodiversity Information Facility (GBIF, www.gbif.org) (Fig. 1), which was carefully validated by an expert on Amazonian avian distribution patterns (author's acronym, L.N.). For each locality, L.N. classified all species by habitat type(s) based on his field experience, complemented by published sources. Bird species lists and habitat classification are available as Supplementary material (Table S4).

Statistical analyses. Since the number of observed OTUs was dependent on the number of reads, we first rarefied all samples to the lowest number of reads obtained from any one plot (23,132 for 16S, 25,144 for 18S and 25,280 for COI; Fig. S1) using the function "rarefy" of the R package vegan v. 2.5–4[83]. For 18S, we discarded one sample ("SJAUTFP1") with a very low number of reads (1,395). As the rarefaction and richness estimates could be biased by rare OTUs[89], we also calculated OTU diversity of order q = 1, which is equivalent to the exponential of the Shannon entropy[90]. We did so by transforming the read counts using the "varianceStabilizingTransformation" function in DESeq. 2[91] as suggested by McMurdie & Holmes[92]. This transformation normalizes the count data with respect to sample size (number of reads in each sample) and variances, based on fitted dispersion-mean relationships[91]. As the results were virtually identical (Pearson correlation > 0.99 for all data sets) we used the

richness based on rarefaction of OTUs for further analyses, since we had no abundance data for birds and just richness measurements was possible. The results of both richness by rarefaction and Shannon estimated are presented in Table S2. As we had three plots in each environment at each locality, we used the mean of the three plots for each environment at each locality.

We tested the relationship between the mean species richness per habitat type of trees and birds by fitting a generalized linear mixed effects model in a Bayesian framework, using Markov chain Monte Carlo (MCMC) methods implemented in the R package MCMCglmm v.2.28[93]. We used this method to control for nested sampling[94], because our plots are nested in the habitat types and we pooled all of them into one regression, but they might differ in their intercept. In this case a mixed effects model would be better suited, since it allows different intercepts. To test the relationship between OTU richness and species richness, we also fitted generalized linear mixed effects models using the OTU richness as the response variable and the genetic marker (16S, 18S and COI), sample type (soil, litter and insects) and tree or bird richness as explanatory variables. We used the Poisson family distribution in the model and considered locality and habitat type as random effects in both analyses. Because the organisms' body size[95] and/or the taxonomic reponses to environmental conditions[10] could affect the diversity patterns, we also divided our data into 16S that comprises mostly bacteria and divided our 18S and COI data between protists, fungi and metazoan, and fitted generalized linear mixed effects models separately for each data sets.

To further assess whether there was any tendency for a positive or negative relationship between OTU and taxonomic diversity, we fitted separate generalized linear models between each OTU richness variable (3 markers and 3 sample types, totaling 9 response variables) against the tree and bird richness separately. We assessed the relationship of these variables based on a two-tailed binomial distribution only focusing on the sign of the relationship. The null expectation is that ~50% of all relationships would be positive and ~50% would be negative if there were no underlying patterns and the relationships were independent of each other. An overabundance of either positive or negative relationships can therefore be seen as a significant deviation from the null-expectation. In our analyses, we carried out a total of nine tests (OTU richness for 3 markers and 3 sample types). The combined probability of achieving 0, 1, 8 or 9 positive outcomes out of nine attempts if both positive and negative relationships are equally likely is 0.039. We therefore considered a relationship where 0, 1, 8 or 9 of the slopes were positive as significant.

Permit(s). Collection permits for this study were granted by the Brazilian authorities ICMBio (registration number 48185–2) and IBAMA (registration number 127341). The SisGen registration number is A8A9AB7.

Received: 27 August 2019; Accepted: 25 November 2019;
Published online: 16 December 2019

References

1. Sutherland, W. J. *et al.* Identification of 100 fundamental ecological questions. *J. Ecol.* **101**, 58–67 (2013).
2. Bebber, D. P., Marriott, F. H. C., Gaston, K. J., Harris, S. A. & Scotland, R. W. Predicting unknown species numbers using discovery curves. *Proc. R. Soc. B Biol. Sci.* **274**, 1651–1658 (2007).
3. Chapman, A. D. Numbers of living species in Australia and the world. (2009).
4. Bini, L. M., Diniz-Filho, J. A. F., Rangel, T. F., Bastos, R. P. & Pinto, M. P. Challenging Wallacean and Linnean shortfalls: knowledge gradients and conservation planning in a biodiversity hotspot. *Divers. Distrib.* **12**, 475–482 (2006).
5. Campbell, G., Kuehl, H., Diarrassouba, A., N'Goran, P. K. & Boesch, C. Long-term research sites as refugia for threatened and over-harvested species. *Biol. Lett.* **7**, 723–726 (2011).
6. Taberlet, P., Coissac, E., Pompanon, F., Brochmann, C. & Willerslev, E. Towards next-generation biodiversity assessment using DNA metabarcoding. *Mol. Ecol.* **21**, 2045–2050 (2012).
7. Leray, M. & Knowlton, N. DNA barcoding and metabarcoding of standardized samples reveal patterns of marine benthic diversity. *Proc. Natl. Acad. Sci.* **112**, 2076–2081 (2015).
8. Stat, M. *et al.* Ecosystem biomonitoring with eDNA: metabarcoding across the tree of life in a tropical marine environment. *Sci. Rep.* **7**, 12240 (2017).
9. Cameron, E. K. *et al.* Global mismatches in aboveground and belowground biodiversity. *Conserv. Biol.* **0**, 1–6 (2019).
10. Bahram, M. *et al.* Structure and function of the global topsoil microbiome. *Nature* **560**, 233–237 (2018).
11. Tedersoo, L. *et al.* Global diversity and geography of soil fungi. *Science* **346**(6213), 1256688 (2014).
12. Lentendu, G. *et al.* Consistent patterns of high alpha and low beta diversity in tropical parasitic and free-living protists. *Mol. Ecol.* **27**, 2846–2857 (2018).
13. Nottingham, A. T. *et al.* Microbes follow Humboldt: temperature drives plant and soil microbial diversity patterns from the Amazon to the Andes. *Ecology* **99**, 2455–2466 (2018).
14. Mora, C., Tittensor, D. P., Adl, S., Simpson, A. G. B. & Worm, B. How many species are there on earth and in the ocean? *PLoS Biol.* **9**, 1–8 (2011).
15. Dominati, E., Patterson, M. & Mackay, A. A framework for classifying and quantifying the natural capital and ecosystem services of soils. *Ecol. Econ.* **69**, 1858–1868 (2010).
16. Zilber-Rosenberg, I. & Rosenberg, E. Role of microorganisms in the evolution of animals and plants: the hologenome theory of evolution. *FEMS Microbiol. Rev.* **32**, 723–735 (2008).
17. O'Malley, M. A. & Dupré, J. *Size doesn't matter: Towards a more inclusive philosophy of biology*. *Biology and Philosophy* **22**, (2007).
18. Hoorn, C. *et al.* Amazonia Through Time: Andean. *Science (80-.).* **330**, 927–931 (2010).
19. Rangel, T. F. *et al.* Modeling the ecology and evolution of biodiversity: biogeographical cradles, museums, and graves. *Science (80-.).* **361**, eaar5452 (2018).
20. Zizka, A., Steege, H., ter, Pessoa, M., do, C. R. & Antonelli, A. Finding needles in the haystack: where to look for rare species in the American tropics. *Ecography (Cop.).* **41**, 321–330 (2018).
21. ter Steege, H. T. *et al.* A spatial model of tree alpha-diversity and tree density for the Amazon. *Biodivers. Conserv.* **12**, 2255–2277 (2003).
22. Bass, D. & Cavalier-Smith, T. Phylum-specific environmental DNA analysis reveals remarkably high global biodiversity of Cercozoa (Protozoa). *Int. J. Syst. Evol. Microbiol.* **54**, 2393–2404 (2004).
23. Bates, J. M. Avian diversification in Amazonia: evidence for historical complexity and a vicariance model for a basic diversification pattern. *Divers. biológica e Cult. da Amaz.* 119–137 (2001).

24. Lovejoy, N. R., Albert, J. S. & Crampton, W. G. R. Miocene marine incursions and marine/freshwater transitions: Evidence from Neotropical fishes. *J. South Am. Earth Sci.* **21**, 5–13 (2006).
25. Antonelli, A., Nylander, J. A. A., Persson, C. & Sanmartín, I. Tracing the impact of the Andean uplift on Neotropical plant evolution. *Proc. Natl. Acad. Sci.* **106**, 9749–9754 (2009).
26. Tuomisto, H. *et al.* Effect of sampling grain on patterns of species richness and turnover in Amazonian forests. *Ecography (Cop.).* **40**, 840–852 (2017).
27. Ter Steege, H. *et al.* Continental-scale patterns of canopy tree composition and function across Amazonia. *Nature* **443**, 444–447 (2006).
28. Silva, S. M. *et al.* A dynamic continental moisture gradient drove Amazonian bird diversification. *Sci. Adv.* **5**, eaat5752 (2019).
29. ter Steege, H. & Hammond, D. S. Character convergence, diversity, and disturbance in tropical rain forest in Guyana. *Ecology* **82**, 3197–3212 (2001).
30. Haugaasen, T. & Peres, C. A. Floristic, edaphic and structural characteristics of flooded and unflooded forests in the lower Rio Purús region of central Amazonia, Brazil. *Acta Amaz.* **36**, 25–35 (2006).
31. Myster, R. W. The physical structure of forests in the Amazon Basin: a review. *Bot. Rev.* **82**, 407–427 (2016).
32. Assis, R. L. *et al.* Patterns of tree diversity and composition in Amazonian floodplain paleo-várzea forest. *J. Veg. Sci.* **26**, 312–322 (2015).
33. Borges, S. H. *et al.* Birds of Jaú National Park, Brazilian Amazon: species check-list, biogeography and conservation. *Ornitol. Neotrop.* **12**, 109–140 (2001).
34. Fine, P. V. A., García-Villacorta, R., Pitman, N. C. A., Mesones, I. & Kembel, S. W. A floristic study of the white-sand forests of Peru. *Ann. Missouri Bot. Gard.* 283–305 (2010).
35. Stropp, J., Van Der Sleen, P., Assunção, P. A., da SILVA, A. L. & ter Steege, H. Tree communities of white-sand and terra-firme forests of the upper Rio Negro. *Acta Amaz.* **41**, (2011).
36. Draper, F. C. *et al.* Peatland forests are the least diverse tree communities documented in Amazonia, but contribute to high regional beta-diversity. *Ecography (Cop.).* **41**, 1256–1269 (2018).
37. Bueno, G. T., Cherem, L. F. S., Toni, F., Guimarães, F. S. & Bayer, M. Amazonia. In *The Physical Geography of Brazil* 169–197 (Springer, (2019).
38. Ter Steege, H. *et al.* An analysis of the floristic composition and diversity of Amazonian forests including those of the Guiana Shield. *J. Trop. Ecol.* **16**, 801–828 (2000).
39. Ter Steege, H. *et al.* Rarity of monodominance in hyperdiverse Amazonian forests. *Sci. Rep.* **9**, 1–15 (2019).
40. Ritter, C. D. *et al.* Locality or habitat? Exploring predictors of biodiversity in Amazonia. *Ecography.* **42**, 321–333 (2019).
41. Vasco-Palacios, A. M., Hernandez, J., Peñuela-Mora, M. C., Franco-Molano, A. E. & Boekhout, T. Ectomycorrhizal fungi diversity in a white sand forest in western Amazonia. *Fungal Ecol.* **31**, 9–18 (2018).
42. Blaxter, M. *et al.* Defining operational taxonomic units using DNA barcode data. 1935–1943 https://doi.org/10.1098/rstb.2005.1725 (2005).
43. Ritter, C. D. *et al.* Biodiversity assessments in the 21st century: The potential of insect traps to complement environmental samples for estimating eukaryotic and prokaryotic diversity using high-throughput DNA metabarcoding. *Genome* **62**, 147–159 (2019).
44. Deiner, K. *et al.* Environmental DNA metabarcoding: Transforming how we survey animal and plant communities. *Mol. Ecol.* **26**, 5872–5895 (2017).
45. Thomsen, P. F. & Willerslev, E. Environmental DNA–An emerging tool in conservation for monitoring past and present biodiversity. *Biol. Conserv.* **183**, 4–18 (2015).
46. Pereira, E. J. *et al.* Policy in Brazil (2016–2019) threaten conservation of the Amazon rainforest. *Environ. Sci. Policy* **100**, 8–12 (2019).
47. Ricklefs, R. E. Community diversity: relative roles of local and regional processes. *Science (80-.).* **235**, 167–171 (1987).
48. Witman, J. D., Etter, R. J. & Smith, F. The relationship between regional and local species diversity in marine benthic communities: a global perspective. *Proc. Natl. Acad. Sci.* **101**, 15664–15669 (2004).
49. da C Jesus, E., Marsh, T. L., Tiedje, J. M. & de S Moreira, F. M. Changes in land use alter the structure of bacterial communities in Western Amazon soils. *ISME J.* **3**, 1004 (2009).
50. Tripathi, B. M. *et al.* Tropical soil bacterial communities in Malaysia: pH dominates in the equatorial tropics too. *Microb. Ecol.* **64**, 474–484 (2012).
51. Rodrigues, J. L. M. *et al.* Conversion of the Amazon rainforest to agriculture results in biotic homogenization of soil bacterial communities. *Proc. Natl. Acad. Sci.* **110**, 988–993 (2013).
52. Mendes, L. W., de L Brossi, M. J., Kuramae, E. E. & Tsai, S. M. Land-use system shapes soil bacterial communities in Southeastern Amazon region. *Appl. soil Ecol.* **95**, 151–160 (2015).
53. de Carvalho, T. S. *et al.* Land use intensification in the humid tropics increased both alpha and beta diversity of soil bacteria. *Ecology* **97**, 2760–2771 (2016).
54. Mueller, R. C. *et al.* Links between plant and fungal communities across a deforestation chronosequence in the Amazon rainforest. *ISME J.* **8**, 1548–1550 (2014).
55. Perner, J. *et al.* Effects of plant diversity, plant productivity and habitat parameters on arthropod abundance in montane European grasslands. *Ecography (Cop.).* **28**, 429–442 (2005).
56. Wenninger, E. J. & Inouye, R. S. Insect community response to plant diversity and productivity in a sagebrush–steppe ecosystem. *J. Arid Environ.* **72**, 24–33 (2008).
57. Bass, M. S. *et al.* Global Conservation Significance of Ecuador' s Yasunı National Park. **5** (1), p.e8767 (2010).
58. Jenkins, C. N., Pimm, S. L. & Joppa, L. N. Global patterns of terrestrial vertebrate diversity and conservation. *Proc. Natl. Acad. Sci.* **110**, E2602–E2610 (2013).
59. Moran, E. F. *et al.* Effects of soil fertility and land-use on forest succession in Amazonia. *For. Ecol. Manage.* **139**, 93–108 (2000).
60. Adeney, J. M., Christensen, N. L., Vicentini, A. & Cohn-Haft, M. White-sand ecosystems in Amazonia. *Biotropica* **48**, 7–23 (2016).
61. Macedo, M. & Prance, G. T. Notes on the vegetation of Amazonia II. The dispersal of plants in Amazonian white sand campinas: the campinas as functional islands. *Brittonia* **30**, 203–215 (1978).
62. Prance, G. T. Islands in Amazonia. *Philos. Trans. R. Soc. London. Ser. B Biol. Sci.* **351**, 823–833 (1996).
63. Edwards, M. E. *et al.* Metabarcoding of modern soil DNA gives a highly local vegetation signal in Svalbard tundra. *The Holocene* **28**, 2006–2016 (2018).
64. Ritter, C. D. *et al.* High-throughput metabarcoding reveals the effect of physicochemical soil properties on soil and litter biodiversity and community turnover across Amazonia. *PeerJ* **6**, e5661 (2018).
65. Kuske, C. R. *et al.* Prospects and challenges for fungal metatranscriptomics of complex communities. *fungal Ecol.* **14**, 133–137 (2015).
66. Urbanová, M., Šnajdr, J. & Baldrian, P. Composition of fungal and bacterial communities in forest litter and soil is largely determined by dominant trees. *Soil Biol. Biochem.* **84**, 53–64 (2015).
67. Mahé, F. *et al.* Parasites dominate hyperdiverse soil protist communities in Neotropical rainforests. *Nat. Ecol. Evol.* **1**, 1–8 (2017).
68. Wood, S. A. *et al.* Consequences of tropical forest conversion to oil palm on soil bacterial community and network structure. *Soil Biol. Biochem.* **112**, 258–268 (2017).
69. Wardle, D. A. *et al.* Ecological linkages between aboveground and belowground biota. *Science (80-.).* **304**, 1629–1633 (2004).
70. Koh, L. P. *et al.* Species coextinctions and the biodiversity crisis. *Science (80-.).* **305**, 1632–1634 (2004).

71. Andresen, E., Arroyo-Rodríguez, V. & Escobar, F. Tropical biodiversity: The importance of biotic interactions for its origin, maintenance, function, and conservation. In *Ecological networks in the tropics* 1–13 (Springer, (2018).
72. Aljanabi, S. M. & Martinez, I. Universal and rapid salt-extraction of high quality genomic DNA for PCR-based techniques. *Nucleic Acids Res.* **25**, 4692–4693 (1997).
73. Guardiola, M. *et al.* Deep-sea, deep-sequencing: metabarcoding extracellular DNA from sediments of marine canyons. *PLoS One* **10**, e0139633 (2015).
74. Klindworth, A. *et al.* Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Res.* **41**, e1 (2013).
75. Leray, M. *et al.* A new versatile primer set targeting a short fragment of the mitochondrial COI region for metabarcoding metazoan diversity: application for characterizing coral reef fish gut contents. *Front. Zool.* **10**, 34 (2013).
76. Meyer, C. P. Molecular systematics of cowries (Gastropoda: Cypraeidae) and diversification patterns in the tropics. *Biol. J. Linn. Soc.* **79**, 401–459 (2003).
77. Zinger, L. *et al.* DNA metabarcoding—Need for robust experimental designs to draw sound ecological conclusions. *Mol. Ecol.* **28**, 1857–1862 (2019).
78. Edgar, R. C. UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nat. Methods* **10**, 996–998 (2013).
79. Quast, C. *et al.* The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* **41**, D590–D596 (2012).
80. Pruesse, E., Peplies, J. & Glöckner, F. O. SINA: accurate high-throughput multiple sequence alignment of ribosomal RNA genes. *Bioinformatics* **28**, 1823–1829 (2012).
81. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
82. Benson, D. A. *et al.* GenBank. *Nucleic Acids Res.* **46**, D41–D47 (2018).
83. Oksanen, J. *et al.* Vegan: community ecology package. R package version 1.17-4. http://cran.r-project.org. Acesso em 23, 2010 (2010).
84. R Core Team. The R development core team. *R: A Language and Environment for Statistical Computing* **1**, (2003).
85. Guevara, J. E. *et al.* Low phylogenetic beta diversity and geographic neo-endemism in Amazonian white-sand forests. *Biotropica* **48**, 34–46 (2016).
86. Borges, S. H. Bird species distribution in a complex Amazonian landscape: species diversity, compositional variability and biotic–environmental relationships. *Stud. Neotrop. fauna Environ.* **48**, 106–118 (2013).
87. Valente, R. de M. Padrões espaciais em comunidades de aves amazônicas. (2006).
88. Cohn-Haft, M., Whittaker, A. & Stouffer, P. C. A new look at the" species-poor" central Amazon: the avifauna north of Manaus, Brazil. *Ornithol. Monogr.* 205–235 (1997).
89. Haegeman, B. *et al.* Robust estimation of microbial diversity in theory and in practice. *ISME J.* **7**, 1092 (2013).
90. Jost, L. Entropy and diversity. *Oikos* **113**, 363–375 (2006).
91. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq. 2. *Genome Biol.* **15**, 550 (2014).
92. McMurdie, P. J. & Holmes, S. Waste not, want not: why rarefying microbiome data is inadmissible. *PLoS Comput. Biol.* **10**, e1003531 (2014).
93. Hadfield, J. D. MCMC methods for multi-response generalized linear mixed models: the MCMCglmm R package. *J. Stat. Softw.* **33**, 1–22 (2010).
94. Markowetz, F., Kostka, D., Troyanskaya, O. G. & Spang, R. Nested effects models for high-dimensional phenotyping screens. *Bioinformatics* **23**, i305–i312 (2007).
95. Zinger, L. *et al.* Body size determines soil community assembly in a tropical forest. *Mol. Ecol.* **28**, 528–543 (2019).
96. Team, Q. D. QGIS geographic information system. *Open Source Geospatial Found. Proj. Versão* **2**, (2015).

Acknowledgements
We thank Rhian Smith and two anonymous reviewers for valuable comments to the manuscript.We thank the Brazilian authorities ICMBio (registration number 48185-2) and IBAMA (registration number 127341) for the collection permits granted for this research; Anna Ansebo, Sven Toresson and Ylva Heed for laboratory and administrative assistance; and Mats Töpel for help with bioinformatics. We thank all plot owners of the Amazon Tree Diversity Network who contributed plot data to the ter Steege et al. (2013) publication for allowing us to use unpublished data. The authors acknowledge financial support from Alexander von Humboldt Foundation and CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico - Brazil: 249064/2013-8) for CDR, the Swedish Research Council (B0569601), the European Research Council under the European Union's Seventh Framework Programme (FP/2007-2013, ERC Grant Agreement n. 331024), the Swedish Foundation for Strategic Research, the Biodiversity and Ecosystems in a Changing Climate (BECC) programme for AA. Open access funding provided by University of Gothenburg.

Author contributions
A.A., C.D.R., and S.F. designed the study; C.D.R. conducted fieldwork; C.D.R. led all analyses with contributions from R.H.N., S.F., D.J.B., and A.Z.; C.D.R. conducted lab work with help from Q.H.; H.t.S. provided the tree data; C.D.R and L.N. compiled and L.N. verified the bird data; C.D.R. led the writing of the manuscript with contributions from all authors.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41598-019-55490-3.

Correspondence and requests for materials should be addressed to C.D.R.

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.

www.nature.com/scientificreports/

5 Conclusion

Thesis summary

The goal of my PhD was to perform empirical adaptation genomic investigations using the outstanding threespine stickleback fish on two ecological axes: basic versus acidic habitats in Scotland and lake versus stream in Vancouver Island, Canada by taking the advantage of powerful genomic tools.

First, we aimed to understand the genomic basis of parallel adaptation. Based on the analysis of several populations inhabiting acidic and basic lakes, we found that the parallel adaptation observed at the ecological and phenotypic level was mirrored at the genomic level with the identification of multiple genomic regions selected differentially between those two ecotypes. In addition, we highlighted that adaptive divergence between acidic and basic stickleback ecotypes occurred via the sorting of pre-existing genetic variation in the marine ancestor (i.e. standing genetic variation) and was predictable from the difference of those derived habitats (basic or acidic) from the ancestral one (marine). These findings led to investigate the process (selective neutrality vs. gene flow) by which standing genetic variation is maintained in the ancestral marine habitat over geographical distances. Using the same acidic and marine populations from Scotland and several new marine populations across the Atlantic, we found that variants selected in derived habitats persist not (only) because of gene flow between derived and ancestral populations but at low to moderate frequency without being deleterious in the ancestor. Further genomic investigations in other organismal ancestral populations would be important to see if these outcomes can be uncovered at global scale.

Second, we investigated how a strong adaptive divergence between parapatric lake-stream stickleback in Vancouver Island, Canada can promote reproductive isolation. We performed, for the first time in that particular stickleback system, a clinal analysis at a fine geographic scale along the habitat transition using the power of whole-genome sequencing. Our results demonstrated the maintenance of reproductive isolation by polygenic adaptation, which constitutes a genome-wide barrier to gene flow without physical isolation.

Overall, the work performed in this thesis confirmed the importance of standing genetic variation and the polygenic nature of adaptation in the stickleback fish and provided solid bases to further explore the two systems from Scotland and Canada investigated during those four years. However, it remains difficult with the methodology described in this thesis to identify with certainty genes being involved in those adaptations. Indeed, a SNP under high genetic differentiation between two ecotypes is not necessarily located on a precise gene as it can also be located several kilobases away on a regulatory element of a gene. In that case, when working only with genome scans, as we did, only assumptions can be made for possible candidate genes and this is not ideal. Further work in that direction would imply genome-wide association mapping to link precise phenotypes (for example the acidic phenotype of North Uist lakes) to a set of genes that could be validated by CRISPR-Cas9 (Clustered Regularly Interspaced Short Palindromic Repeats) editing and the study of regulatory evolution to untangle more precisely which genes are driving the phenotypic variability we observe between stickleback ecotypes and at which stage they are expressed during the development.

On a more personal point of view...

When I first arrived at the Zoological Institute, I gave a talk to present myself and titled it “I am going on an adventure”, quoting The Hobbit. And this PhD was definitely an adventure! Not an easy one but a great one nonetheless! It has strengthened my abilities in molecular bench work and bioinformatics and I found out what I was capable of and learnt a lot about myself.