

Genomics of Adaptation Revealed in Threespine Stickleback

Inauguraldissertation
zur
Erlangung der Würde eines Doktors der Philosophie
vorgelegt der
Philosophisch-Naturwissenschaftlichen Fakultät
der Universität Basel

von Marius Roesti
aus Kandersteg (BE)

Basel, 2016

[Originaldokument gespeichert auf dem Dokumentenserver der Universität Basel](#)

edoc.unibas.ch



[Dieses Werk ist lizenziert unter einer Creative Commons Namensnennung 4.0.](#)

Genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultät auf Antrag von
Prof. Dr. Walter Salzburger und Prof. Dr. Roger Butlin

Basel, 9. Dezember 2014

Dekan Prof. Dr. Jörg Schibler

Dissertation

Genomics of Adaptation Revealed in Threespine Stickleback

Marius Roesti



Zoologisches Institut
University of Basel
2014

Supervised by Prof. Dr. Walter Salzburger & Dr. Daniel Berner
Second examiner: Prof. Dr. Roger Butlin

Preface

“As many more individuals of each species are born than can possibly survive; and as, consequently, there is a frequently recurring struggle for existence, it follows that any being, if it vary however slightly in any manner profitable to itself, under the complex and sometimes varying conditions of life, will have a better chance of surviving, and thus be NATURALLY SELECTED. From the strong principle of inheritance, any selected variety will tend to propagate its new and modified form.”

Charles Darwin, 1859.

(On the Origin of Species by Means of Natural Selection, p.5, 1st edition)

Darwin revolutionized how we see the world and our place therein. In 1859, 23 years after returning with the H.M.S. Beagle from the Galapagos, he published his idea¹ of evolutionary change through natural selection on more than 400 pages. In fact, Darwin regarded this work only a summary of all the evidence he had gathered. *On the Origin of Species by Means of Natural Selection* caused an outrage in the civilized world. The book coherently describes ubiquitous natural principles derived from meticulously reflected observations, which also changed the view on the origin of our own species.

In the first half of the 20th century, Darwin’s evolutionary principles were integrated with early insights in genetics. This ‘modern evolutionary synthesis’ is regarded as one of the major leaps (evolutionary) biology has taken since the Darwinian revolution. In particular, this synthesis emphasized biology as an integrative, yet independent science with a pervasive evolutionary thinking. The so far observational principles of evolutionary change through natural selection had become, to a certain degree, testable and measurable.

Darwin’s principles also withstand biology’s current technical revolution. Clearly, I am part of this modern era of ‘technically enabled biologists’. My every day’s work and my insights are much determined by the availability and use of powerful technologies. In the current scientific world, it is easy to forget about past accomplishments, which have brought us technologically as well as ideologically to where we are today. We are standing on the shoulders of giants allowing us to see further than we have ever seen before (Newton 1676).

I am often confronted with the question ‘why do you do science?’. To me, being a scientist means to be an explorer, an explorer who is insatiably curious about what might or might not be, and ideally, who has the freedom to be creative in the way he reaches new grounds for a better understanding of the world. “To make a contribution to this better understanding”, as Ernst Mayr (1998) puts it, “is a source of great

satisfaction to a scientist; indeed it is an occasion for exhilaration. The emphasis is often on discovery, where luck sometimes plays a role, but the joy is perhaps even greater when one succeeds in the difficult intellectual achievement of developing a new concept, a concept that can integrate a mass of previously disparate facts, or one that is more successful as the basis of scientific theories.”

Another important question to me is ‘why biology?’. I feel deep joy when I explore and ponder upon the living world with all its so diverse beauty and hidden principles. Even though my scientific explorations into this living world have inevitably a relatively specific focus, they make only sense in context of the ‘bigger picture’, to which they hopefully add. And, it is likely the (non-scientific) fascination for nature I have since I remember (and which I can hardly put into words) that brought me to finally become a biologist.

Marius Roesti, November 2014

¹ Alfred Russel Wallace should be credited with co-discovering, together with Darwin, the principles of evolutionary change through natural selection. Even though it seems that Darwin had gathered a more complete and larger body of evidence, the first ever publication on natural selection appeared as a co-authored manuscript by Darwin and Wallace (1858).

References

Darwin (1859) *On the Origin of Species by Means of Natural Selection*, London.

Darwin C, Wallace AR (1858) On the Tendency of Species to Form Varieties; and on the Perpetuation of Varieties and Species by Natural Means of Selection. *Zoological Journal of the Linnean Society* 3, 46-50.

Mayr E (1997) *This is Biology: The Science of the Living World*. Harvard University Press.

Newton I (1676) Letter to Robert Hooke.

Contents

Preface	7
Thesis Summary	13
Thesis Introduction	17
Main Chapters	
Chapter 1: Genome divergence during evolutionary diversification as revealed in lake-stream stickleback population pairs.....	27
Supplementary material.....	43
Chapter 2: Uninformative polymorphisms bias genome scans for signatures of selection.....	55
Chapter 3: Recombination in the threespine stickleback genome—patterns and consequences.....	65
Supplementary material.....	83
Chapter 4: The genomic signature of parallel adaptation from shared genetic variation.....	89
Supplementary material.....	107
Chapter 5: Genetic architecture of skeletal evolution in European lake and stream stickleback.....	141
Supplementary material.....	159
Chapter 6: The genomics of ecological vicariance in threespine stickleback fish.....	167
Supplementary material.....	187
Outreach Chapters	
Chapter 7: Strong genome-wide divergence between sympatric European river and brook lampreys.....	213
Supplementary material.....	219
Chapter 8: Natural selection: it's a many-small world after all.....	231
Chapter 9: Field work	
Chapter 9.1: Repeated lake-stream divergence in stickleback life history within a central European lake basin.....	239
Supplementary material.....	253
Chapter 9.2: Depth-dependent abundance of Midas Cichlid fish (<i>Amphilophus</i> spp.) in two Nicaraguan crater lakes.....	259
Supplementary material.....	271
Chapter 9.3: Testing the stages model in the adaptive radiation of cichlid fishes in East African Lake Tanganyika.....	277
Supplementary material.....	291
Thesis Conclusion	307
Acknowledgements	313

Thesis Summary

Natural selection is the ultimate, but not only force underlying organismal diversity. Despite this general biological insight, our understanding of how selection targets and shapes the genome during adaptation remains incomplete and is the central quest of this thesis. My main model organism is the threespine stickleback fish (*Gasterosteus aculeatus*). Stickleback provide an outstanding opportunity to study adaptive evolutionary change, because marine ancestors have repeatedly colonized and adapted to different freshwater environments all over the northern hemisphere since the last glacial retreat about 12,000 years ago. Besides wild populations, I also make use of lab-raised stickleback hybrids from controlled crosses for this thesis work.

Thousands of genome-wide genetic polymorphisms (i.e., genetic markers) called in marine, but predominantly in distinct lake and stream stickleback populations from different geographic locations allow me to decipher the number and position of genomic targets of selection in the early phase of adaptive divergence. I find that selection acts on many loci distributed widely across the genome. On a genomic scale, the recombination landscape along chromosomes proves to be – in concert with selection – an important factor in driving heterogeneous genetic differentiation among populations.

To investigate the rate of recombination across the stickleback's genome in more detail, I use an artificially crossed second-generation (F2) population. This reveals constraints in the frequency and location of detectable recombination events (i.e., cross-overs) within the genome. For example, cross-overs prove to be more frequent in chromosome peripheries than centers. This, together with selection, results in decreased within-population genetic diversity and increased between-population differentiation in the centers of chromosomes as opposed to the peripheries. Furthermore, I show that the cessation of recombination between the heterogametic sex chromosomes occurred in independent bouts. As a consequence, I find extended genomic regions distinct in their degree of degeneration between the X and Y chromosome, so called evolutionary strata. Finally, recombination reveals to be an important determinant of other aspects of a genome, such as its nucleotide composition.

Integrating theoretical modeling with targeted and genome-wide sequencing, my research further demonstrates that the inference and interpretation of genomic regions exhibiting particularly high and low population differentiation is not as straightforward as commonly believed. This is because the type of genetic variation available to selection (i.e., pre-existing vs. de novo variation) as well as the mode of adaptation (i.e., divergent vs. parallel adaptation) influence the way neutral variation is shaped by selection across the genome. I demonstrate that a genomic region of high differentiation may not necessarily be indicative of divergent selection when populations adapt in parallel to similar environments from a shared pool of genetic variation.

Based on several hundreds of F2 specimens reared under standardized conditions in the laboratory, I also link variation in heritable phenotypic traits to genetic variation, a research program generally referred

to as quantitative trait locus (QTL) mapping. Corroborating with the results from my genome scans within and between wild populations (indicating that adaptive divergence involves many loci widespread across the genome), QTL mapping reveals that most phenotypic traits are controlled by numerous genetic loci. In general, each of these loci explains a small fraction of the entire phenotypic trait variation.

I also use high resolution SNP data to infer the demographic history of several lake and stream stickleback populations from the Lake Constance watershed (Central Europe) and demonstrate that the repeated occurrence of similar stream phenotypes are, in this particular system, better explained by an evolutionary scenario of 'ecological vicariance' rather than repeated parallel divergence. I then show how selection has shaped local and broad-scale linkage, diversity and differentiation across the genome in these populations. Interestingly, I find evidence for strong divergent selection acting on large chromosomal rearrangements I had previously detected to be important for marine vs. freshwater adaptation. This finding provides a strong case for the re-use of pre-existing genetic variation in stickleback and demonstrates that the same genomic regions can be involved in adaptive divergence between disparate ecotype pairs.

Overall, I come to conclude that signatures of selection are – at various physical scales – frequent within the stickleback genome. Stickleback repeatedly use pre-existing genetic variation, shared across various geographic ranges, to adapt to similar or disparate environments. Yet, there is a substantial degree of genetic non-parallelism – at least at the level of neutral markers – that goes along with parallel phenotypic evolution. My thesis emphasizes that the reliable detection and interpretation of genomic signatures of selection requires integrating many replicate study populations within a clear-cut ecological and demographic framework, as well as complementary analytical approaches. Controlled crossing experiments and theoretical modeling are key to deriving predictions about the genomics of adaptation in the wild and to facilitate our understanding of complex biological processes and patterns.

Thesis Introduction

(i) Adaptation and Speciation Research

Adaptation is the evolutionary response of an organism to fit its proximate environment. A locally adapted population should thus perform best in its native environment and become distinct from any other population occupying a foreign environment (Kawecki & Ebert 2005). Ultimately, this process can lead to the build-up of reproductive isolation between initially interbreeding populations, to which we then refer to as distinct 'species' (Schluter 2000, Nosil 2012). – Even though Charles Darwin and Alfred Russell Wallace laid the basis of this ecological (adaptive) idea of speciation more than 150 years ago (Darwin & Wallace 1858; Darwin 1859), a non-ecological view of the speciation process dominated evolutionary biology over the majority of the last century (Mayr 1997). For a long time, spatial isolation was considered essential for population divergence and speciation. However, the relatively recent and rapid accumulation of empirical and theoretical evidence in favor of ecological speciation demonstrates that ecology only – that is, in the absence of geographical barriers, can reduce gene flow between populations. In fact, ecological speciation appears to be the dominant mode of speciation (Schluter 2000; Rundle & Nosil 2005; Nosil 2012). This emphasizes natural selection's prime role in the origin of organismal diversity.

Speciation research has thus shifted its focus to investigating the relevance of ecology (adaptation) for evolutionary change. Ideally, we examine closely related populations or species that have diverged multiple times independently between ecologically distinct habitats in the recent evolutionary past. This is because (i) we are interested in the factors driving the *initial* phase of adaptive divergence between populations, (ii) the presence (or absence) of *repeated* evolutionary patterns among independent, but ecologically comparable populations provides evidence whether the ultimate force underlying these patterns has been adaptive or rather stochastic, and (iii) relatively recent population divergence provides insights into the rate of adaptive evolution in the wild.

(ii) Evolutionary Genomics and its Questions

For a long time, evolutionary biology has been technically constrained to investigations at the phenotype-level. In the past two decades, however, we have witnessed a revolution in the methodological toolkit available to evolutionary biologists. Enabled by advances in high-throughput nucleic acid sequencing technologies, many of us are currently trying to understand adaptation at the genomic level. The screening of entire genomes, or representative subsets thereof, has become standard in the pursuit of answering some key questions in (evolutionary) biology:

- Is genome divergence between differentiating populations homo- or heterogeneous?

- How many genomic regions are targeted by selection during adaptation, and where are they located in the genome?
- Are the same genetic loci repeatedly involved when populations adapt in parallel to similar environments?
- What type of genetic variation (pre-existing vs. de novo variation) does selection act upon during early adaptation?
- How does selection impact neutral genetic variation within the genome?
- What is the genetic basis of heritable phenotypic variation? How many loci control a phenotypic trait and what is the relative contribution of each of these loci to the overall trait variation?
- What is the role of selection relative to, or in concert with other processes and factors involved in genome evolution, such as the rate of crossing over along the genome, or the demographic history of populations?

Finding answers to the above questions will shed light on the predictability of adaptive evolution in general. To what extent would evolution take the same path, if we replayed the tape of life (Gould 1989)?

(iii) Evolutionary Genomics Approaches

Genome scans in natural populations

Different analytical approaches can be taken to screen the genome of wild populations for signatures of selection (Oleksyk et al. 2010). A promising and now broadly applied approach is 'divergence mapping', which investigates the extent of differentiation between populations inhabiting different environments (Nielsen 2005; Storz 2005). Ideally, we can rely on a well-assembled reference genome of the focal species allowing us to integrate single point estimates of divergence along the genome. The premise of divergence mapping is that natural selection influences the extent of population differentiation around selected loci in contrast to the neutrally evolving remainder of the genome (Wu 2001). Regions exhibiting strong population divergence are thought to hold loci where different variants (alleles) are selectively favored between populations – that is, they are under divergent selection. In this case, increased genetic differentiation should extend beyond the selected locus. This is because not only the locus directly targeted by selection, but also its physically linked genomic neighborhood experiences allele frequency shifts (Maynard Smith & Haigh 1974). Regions of particularly low genetic differentiation, on the other hand, are generally interpreted as either being under balancing selection (i.e., selection favoring genetic diversity), or to represent regions in the genome where lineage sorting between populations is (yet) incomplete (e.g., due to ongoing gene flow). A traditional measure to quantify molecular population divergence is the 'fixation index' (F_{ST}), which ranges from 0 (no differentiation) to 1 (complete differentiation) and is calculated based on allele frequencies within and between populations (Wright 1950; Weir & Cockerham 1984; Holsinger et al. 2009). More recently, and complementary to F_{ST} -like measures, phylogenetic methods have become popular to infer population structure along the genome (e.g., Heliconius Genome Consortium 2012; Jones et al. 2012).

Divergence mapping makes explicit use of comparing populations occupying different habitats. Nevertheless, genomic footprints of selection can also be detected *within* single populations (which potentially inhabit similar habitats). For instance, a local reduction in genetic diversity or an increased extent of linkage disequilibrium (LD) within a population can point to a selected region within a genome (Sabeti et al. 2006; Oleksyk et al. 2010). This is because a positively selected variant is expected to increase, together with its physically linked neutral genomic neighborhood, in frequency, thereby locally reducing genetic variation and increasing the extent of linkage-segments. The problem here is that other processes besides selection can alter diversity and LD patterns within a genome. Most of them, however, should impact the entire genome (e.g., genetic bottlenecks) instead of only localized genomic regions. It further proves difficult, when the focus is on a single population only (or several populations that inhabiting

a similar environment), to link detected regions under selection to a specific ecological context. This in turn can be an advantage of within-population selection measures, as we do not have to make pre-assumptions about an ecological contrast (i.e., rely on divergent selection pressures) among populations. Within-population measures for selection thus provide a strong means for detecting *any kind* of selected regions within a genome. Between- and within-population measures also differ in other aspects, such as in the use of different polymorphisms as genetic markers (possibly resulting in different study resolutions). For example, polymorphisms with a relatively low minor allele frequency (MAF) may be informative when we are interested in investigating diversity or linkage within a genome, but can be uninformative in a divergence mapping approach (Bhatia et al. 2013).

Pedigree data from artificial populations

Besides divergence scans, where we investigate genomes of wild populations, experimental approaches can be powerful to decipher certain aspects important to the genomics of adaptation. Many such experimental genomics approaches are based on controlled crosses in the laboratory, as it is the case for 'quantitative trait locus' (QTL) mapping. The objective here is to link genetic marker data with inherited trait variation in second-generation hybrids from an artificial cross-population raised under standardized laboratory conditions (to control for phenotypic plasticity). This allows connecting focal phenotypic traits to certain genomic regions and thus to uncover the genetic architecture of phenotypic variation selection acts upon in the wild. QTL mapping provides a means to answer questions about the number and location of genomic regions, as well as their relative contributions to trait variation underlying (adaptive) phenotypes. Nevertheless, there are some inherent limitations to QTL mapping, such as the limited genomic mapping resolution (given by a single generation of recombination from the F1 to the F2 generation), or the possibility to have a non-representative subset of allelic variants present in an artificial cross relative to the wild population (Rockman 2010; Savolainen et al. 2013).

Pedigree data, as provided by a QTL hybrid-cross, can further be used to study genomic recombination, that is – the process of genetic exchange between chromosomes during meiosis. Recombination is an important evolutionary mechanism (Butlin 2005). Through the shuffling and breaking of chromosomal DNA segments, it increases genetic diversity and allows selection to operate more efficiently. On the other hand, the suppression of recombination is pivotal for the evolution of heterogametic sex chromosomes (Charlesworth et al. 2005). Based on genetic markers fixed for different alleles between the two parental individuals of a hybrid-cross, we can infer properties about recombination, such as the cross-over rate along chromosomes, when screening the same markers again in the F2 or a later generation.

Association mapping

'Association mapping' provides a solution to some of the above-mentioned limitations given by QTL mapping (Balding 2006; Oleksyk et al. 2010). As in QTL mapping, the goal in association mapping is to find a statistical association between genetic markers and quantitative trait variation. Contrary to QTL mapping, however, association mapping relies on closely related wild individuals (ideally from the same population or a hybrid zone), exhibiting variation in the phenotypic trait(s) of interest. The approach relies on genomic linkage between genetic markers and the actual causative polymorphisms – that is, the actual polymorphisms responsible for the observed phenotypic trait variation. Because LD normally extends on relatively short genomic distances in wild populations (some hundred to a few thousand base pairs), association mapping provides a much higher mapping-resolution than QTL mapping, commonly down to one or few candidate genes. Because association mapping relies on wild individuals, the naturally occurring allelic richness available to selection is much better represented in this approach compared to QTL mapping. Nevertheless, the available number of individuals and genetic markers commonly limit the power of association mapping studies. Furthermore, because the focus is on natural samples, other factors such as population genetic structure or population history can lead to false-positive results.

(iv) Reduced Representation Genome Sequencing

Evolutionary genomics approaches rely on abundant genetic polymorphisms (markers) within a genome. Despite the staggering amount of data we are now able to produce by next-generation sequencing technologies, obtaining DNA sequences of entire genomes for genetic non-model organisms is still prohibitively expensive. This is especially the case when studying organisms with relatively large genomes and when we aim for a sufficiently extensive sample number that is representative of a large natural or experimental population rather than few individuals only. Depending on the study objectives, obtaining sequence data from every nucleotide in the genome can provide more data than necessary. Researchers in the field of population genomics thus often employ reduced representation genome sequencing approaches, which provide relatively high-resolution molecular data across a subset of the genome (Davey et al. 2011; Narum et al. 2013). A now widely applied technique is 'Restriction-site Associated DNA' (RAD) sequencing (Baird et al. 2008). Here, the entire DNA of single individuals is cut by a restriction enzyme at specific and abundant genomic sequence-motifs. After this enzymatic digest, unique barcodes are ligated to the resulting DNA-fragments' sticky ends. In a final step, a next-generation sequencing platform, such as Illumina sequencing, is used to obtain around 150 bp-long sequence-reads from either side of all the genome-wide restriction enzyme cutting sites. Based on the barcode information sequenced together with the genomic DNA, each of the many millions of obtained reads can be assigned back to a specific individual. Overall, RAD sequencing allows us to individually sequence 'representative' parts of a genome at thousands of homologous sites to a relatively high coverage in many individuals (Davey et al. 2011). Furthermore, by choosing different enzymes with different cutting frequencies, we can effectively alter the number of represented sites within a genome.

(v) Evolutionary Model: Threespine Stickleback Fish

The threespine stickleback (*Gasterosteus aculeatus*) is a relatively small teleost fish with a well-understood natural history: marine ancestors have repeatedly invaded freshwater all over the northern hemisphere since the last glaciation period about 12,000 years ago (Münzing 1963; Bell & Foster 1994) (Figure 1). Following this initial freshwater colonization and adaptation, stickleback have – despite the opportunity for gene flow – further diversified between different habitats such as lakes and streams (Figure 1), or the benthic and limnetic niche within single lakes. These features make the threespine stickleback an outstanding model to study the early stages of adaptive divergence in many natural replicates.

Within ecologically similar habitats, distinct phenotypic shifts in morphology, behavior, physiology, and life history have repeatedly evolved in stickleback, suggesting that they are adaptive and thus shaped by natural selection (Figure 1). For example, stream fish generally exhibit fewer or shorter gill rakers (bony structures projecting from the branchial arches used to acquire and retain prey) and are deeper-bodied than lake fish (Berner et al. 2008, 2010). These phenotypes are thought to reflect differences in foraging habitats between lakes and streams: slender bodies and more numerous or longer gill rakers aid in feeding on zooplankton in the open water column of lakes, while deeper bodies and fewer or shorter gill rakers aid in feeding on benthic macroinvertebrates in streams. Yet, different stickleback populations also show ample variation in traits unrelated to foraging. Marine stickleback, for instance, exhibit instead of common fish scales numerous bony plates along their body axes (Bell & Foster 1994) (Figure 1). On the contrary, most freshwater stickleback have a reduced number of these plates. These plate-morph differences are thought to represent adaptations to contrasting predation regimes in these habitats. In fact, shifts in plate morphology are not unique across the marine-freshwater transition, but they are also found between some lake and stream populations of the same watershed (Berner et al. 2010).

Although often reproductively isolated in nature, stickleback from different habitats and locations can still be crossed and kept in the laboratory. Furthermore, there is a wealth of molecular tools available for



Figure 1 The left picture shows juvenile stickleback in their natural stream habitat (Rhine, Switzerland). On the upper right picture a typical marine male stickleback in its breeding color is shown (Vancouver Island, BC, Canada). Note the full set of lateral bony plates along the body. The lower right picture shows two full-grown male stickleback, the one on top from a stream population and the one below from the lake from the Lake Constance watershed. All pictures were taken by M. Roesti.

stickleback, most notably a reference genome assembled to the quality of entire chromosomes (Jones et al. 2012). These features make stickleback, together with all the above-mentioned ecological advantages, particularly well suited to address evolutionary genomics questions.

Taken together, threespine stickleback offer a remarkable opportunity to study adaptation, and insights gained from research on this organism have played a critical role in shaping our understanding of evolution.

(vi) Thesis Outline

In this thesis, I address some fundamental questions in evolutionary genomics (see above). I use a combination of different genomic approaches and apply them to wild as well as experimental populations of threespine stickleback. The thesis is structured in six Main Chapters and three Outreach Chapters followed by an overall discussion of the results obtained from the Main Chapters. I emphasize that all the work I here present comes from collaborations with different people (see Thesis Acknowledgements and study-specific authors contributions sections).

For Chapter 1 (Roesti et al. 2012a), I established and refined the RAD sequencing protocol as well as the bioinformatics pipeline to type genome-wide markers in stickleback. I used divergence mapping to investigate the way the genome differentiates in the course of early adaptive divergence between four pairs of lake and stream stickleback populations from Vancouver Island (BC, Canada). I find population differentiation to be highly heterogeneous along the genome. The study shows that overall genome-wide divergence is reflected in the magnitude of divergence in phenotypes known to be under divergent selection across the population-pairs. Chromosomal centers are generally strongly accentuated in divergence as compared to the peripheries, suggesting an important role of recombination for genome evolution. I argue that accounting for this chromosome-wide effect is important if we are interested in finding the actual genomic *targets* of selection. Generally, the study reveals a great number of strongly differentiated regions widely distributed across the genome between lake and stream stickleback. Even

though this finding indicates that adaptation involves many genetic loci, I also demonstrate that high-divergence genomic regions may be challenging to interpret and may, in fact, sometimes be unrelated to the focal ecological contrast. The insights from this study, presented as Chapter 1, led to three follow-up investigations (Chapters 2-4).

Chapter 2 (Roesti et al. 2012b) explores the commonly overlooked problem of ‘uninformative markers’ (i.e., genetic polymorphisms with a low ‘minor allele frequency’, MAF) in divergence mapping approaches. The reason why low-MAF markers should be discarded is that they represent relatively ‘young’ mutations (or sequencing artifacts) that do not adequately capture the focal processes of divergence mapping at the molecular level (i.e., hitchhiking with a selected locus and drift). I empirically demonstrate the bias of uninformative markers on the inference of overall genomic and more localized genetic differentiation within the genome.

Chapter 3 (Roesti et al. 2013) uses an artificial stickleback cross-population to investigate the genome-wide patterns of crossing over (recombination) and their biological consequences and methodological implications. Even though stickleback offer a powerful model system in evolutionary biology, a detailed analysis of recombination in these fish had been missing to that point. In a first step, the study resulted in the improvement of the threespine stickleback reference genome, which is now widely used as the standard reference for genomics investigations. Overall, the study points to strong functional constraints in the rate and distribution of cross-overs along chromosomes. I demonstrate that heterogeneity in cross-over events along chromosomes results in decreased genetic diversity, but increased genetic differentiation in chromosomal centers as opposed to peripheries. Furthermore, I detect a strong association between cross-over rate and GC nucleotide content within the genome, highlighting the general importance of recombination for genome evolution. Finally, recombination has ceased in different bouts along the sex chromosomes in stickleback. For the first time in a fish species, I provide high-resolution evidence that this process has resulted in distinct ‘evolutionary strata’ (i.e., large genomic regions exhibiting distinct degrees of genetic differentiation) between the heterogametic sex chromosomes.

Chapter 4 (Roesti et al. 2014) is a study that presents theory for parallel adaptation, a process believed to be widespread in nature with ample evidence coming from animals, plants, and bacteria. I focus on the situation where multiple populations have used the same genetic variation to adapt to selectively similar environments. Although a common process, the genomic signature of parallel adaptation from shared genetic variation had not been investigated by theoreticians or empiricists to that point. For this work, I integrated theoretical modeling and empirical work (performed in stickleback). The findings offer novel interpretations of genome scans (i.e., for high and low divergence outlier regions). Furthermore, I introduce a new way on how to calculate phylogenetic separation along the genome in the empirical part of this study.

Chapter 5 (Berner et al. 2014) uses QTL mapping to explore the genetic architecture of (adaptive) phenotypic traits in stickleback. Close to 500 F₂-hybrids from a cross between a phenotypically divergent lake male and stream female stickleback were genotyped at more than 2,000 genome-wide genetic markers. The study specifically focuses on the genetic basis of skeletal traits (i.e., plate number, vertebrae number, gill raker length, snout and head length). Even though I find evidence for two genetic loci that account for a large proportion of the observed trait variation, most QTLs prove to have only small phenotypic effects. To draw the link between the laboratory-raised QTL cross population and the wild, SNP data from the natural source populations of the two parental individuals demonstrate that some SNPs linked to QTLs in the cross also exhibit striking allele frequency differences in the wild, suggesting a causal role of these QTLs in adaptive population divergence.

Chapter 6 (Roesti et al. 2015) presents high-resolution genome scans conducted with European lake and stream populations from the Lake Constance watershed. The study first infers the demographic history of the lake and its adjacent stream populations. I find that multiple, phenotypically similar stream populations are the result of ‘ecological vicariance’ and not parallel (i.e., repeated) evolution. In the light of these demographic insights and by integrating different analytical approaches, I demonstrate the way

selection has shaped the genome between and within the divergent lake and stream populations. In particular, I focus on the extent of selective sweeps and their impact on linkage disequilibrium and genetic diversity. I also use an association mapping approach to reveal the genomic architecture of variation in lateral plate morphology, pointing to a yet unknown genomic region that harbors a promising novel candidate gene controlling this trait. Finally, the study demonstrates the successful detection of large chromosomal arrangements (i.e., inversions), which I find to be involved in lake-stream as well as marine-freshwater adaptation. This provides striking evidence for the re-use of globally shared variation to adapt across different habitat transitions in stickleback.

The Outreach Chapters 7-9 include additional work I was involved in during my time as a PhD student. Chapter 7 is a study that investigates genomic differentiation in lampreys (Mateus et al. 2013). We demonstrate, for the first time, that European river (*Lampetra fluviatilis*) and brook (*Lampetra planeri*) lampreys are strongly differentiated within their genomes and thus comprise well-separated fish species. Chapter 8 is a perspective published on a recent Nature paper (Arnegard et al. 2014). Arnegard et al. used an experimental approach to reveal the genetic basis of habitat divergence between benthic and limnetic stickleback. Chapter 9 includes three studies for all of which I was heavily-involved in field sampling. [I further contributed to the acquisition of phenotype data for Muschick et al. (2014), to the study design and data acquisition for Moser et al. (2012), and to the study design and paper writing for Dittmann & Roesti et al. (2012).] Moser et al. (2012) investigate divergence in life history traits between lake and stream stickleback in the European Lake Constance watershed. Dittmann & Roesti et al. (2012) is a study conducted during a field excursion to Central America, which looks at the depth-dependent abundance of Midas cichlid fish between two crater lakes in Nicaragua. Finally, Muschick et al. (2014) investigate whether the adaptive radiation of cichlid fish in the East African Lake Tanganyika has occurred in distinct stages based on a large set of phenotypic trait data.

References

- Arnegard ME, McGee MD, Matthews B, et al. (2014) Genetics of ecological divergence during speciation. *Nature* 511, 307-311.
- Baird NA, Etter PD, Atwood TS, et al. (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE* 3, e3376.
- Balding DJ (2006) A tutorial on statistical methods for population association studies. *Nature Reviews Genetics* 7, 781-791.
- Bell MA, Foster SA (1994) *The evolutionary biology of the threespine stickleback*. Oxford University, Oxford.
- Berner D, Adams DC, Grandchamp AC, Hendry AP (2008) Natural selection drives patterns of lake-stream divergence in stickleback foraging morphology. *Journal of Evolutionary Biology* 21, 1653-1665.
- Berner D, Moser D, Roesti M, Buescher H, Salzburger W (2014) Genetic architecture of skeletal evolution in European lake and stream stickleback. *Evolution* 68, 1792-1805.
- Berner D, Roesti M, Hendry AP, Salzburger W (2010) Constraints on speciation suggested by comparing lake-stream stickleback divergence across two continents. *Molecular Ecology* 19, 4963-4978.
- Bhatia G, Patterson N, Sankararaman S, Price AL (2013) Estimating and interpreting FST: the impact of rare variants. *Genome Research* 23, 1514-1521.
- Butlin RK (2005) Recombination and speciation. *Molecular Ecology* 14, 2621-2635.
- Charlesworth D, Charlesworth B, Marais G (2005) Steps in the evolution of heteromorphic sex chromosomes. *Heredity* 95, 118-128.
- Darwin C (1859) *On the Origin of Species by Means of Natural Selection*. London.
- Darwin C, Wallace AR (1858) On the Tendency of Species to Form Varieties; and on the Perpetuation of Varieties and Species by Natural Means of Selection. *Zoological Journal of the Linnean Society* 3, 46-50.
- Davey JW, Hohenlohe PA, Etter PD, et al. (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics* 12, 499-510.
- Dittmann MT*, Roesti M*, Indermaur A, et al. (2012) Depth-dependent abundance of Midas Cichlid fish (*Amphilophus* spp.) in two Nicaraguan crater lakes. *Hydrobiologia* 686, 277-285.
- Gould SJ (1989) *Wonderful Life: The Burgess Shale and the Nature of History*. Norton, New York.
- Heliconius Genome Consortium (2012) Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature* 487, 94-98.
- Holsinger KE, Weir BS (2009) Genetics in geographically structured populations: defining, estimating and interpreting Fst. *Nature Reviews Genetics* 10, 639-650.
- Jones FC, Grabherr MG, Chan YF, et al. (2012) The genomic basis of adaptive evolution in threespine sticklebacks. *Nature* 484, 55-61.
- Kawecki TJ, Ebert D (2004) Conceptual issues in local adaptation. *Ecology Letters* 7, 1225-1241.
- Mateus CS, Stange M, Berner D, Roesti M, et al. (2013) Strong genome-wide divergence between sympatric European river and brook lampreys. *Current Biology* 23, R649-650.
- Maynard Smith J, Haigh J (1974) Hitch-hiking effect of a favorable gene. *Genetics Research* 23, 23-35.
- Mayr E (1997) *This is Biology: The Science of the Living World*. Harvard University Press.

- Moser D, Roesti M, Berner D (2012) Repeated lake-stream divergence in stickleback life history within a Central European lake basin. *PLoS ONE* 7, e50620.
- Münzing J (1963) The evolution of variation and distributional patterns in European populations of the three-spined stickleback, *Gasterosteus aculeatus*. *Evolution* 17, 320-332.
- Muschick M, Nosil P, Roesti M, et al. (2014) Testing the stages model in the adaptive radiation of cichlid fishes in East African Lake Tanganyika. *Proceedings of the Royal Society B* 281, 20140605.
- Narum SR, Buerkle CA, Davey JW, Miller MR, Hohenlohe PA (2013) Genotyping-by-sequencing in ecological and conservation genomics. *Molecular Ecology* 22, 2841-2847.
- Nielsen R (2005) Molecular signatures of natural selection. *Annual Review of Genetics* 39, 197-218.
- Nosil P (2012) *Ecological Speciation*, 1 edn. Oxford Series in Ecology and Evolution.
- Oleksyk TK, Smith MW, O'Brien SJ (2010) Genome-wide scans for footprints of natural selection. *Philosophical Transactions of the Royal Society of London Series B* 365, 185-205.
- Rockman MV (2012) The QTN program and the alleles that matter for evolution: all that's gold does not glitter. *Evolution* 66, 1-17.
- Roesti M, Gavrillets S, Hendry AP, Salzburger W, Berner D (2014) The genomic signature of parallel adaptation from shared genetic variation. *Molecular Ecology* 23, 3944-3956.
- Roesti M, Hendry AP, Salzburger W, Berner D (2012a) Genome divergence during evolutionary diversification as revealed in replicate lake-stream stickleback population pairs. *Molecular Ecology* 21, 2852-2862.
- Roesti M, Kueng B, Moser D, Berner D (2015) The genomics of ecological vicariance in threespine stickleback fish. *Nature Communications* 6, doi: 10.1038/ncomms9767.
- Roesti M, Moser D, Berner D (2013) Recombination in the threespine stickleback genome—patterns and consequences. *Molecular Ecology* 22, 3014-3027.
- Roesti M, Salzburger W (2014) Natural Selection: It's a Many-Small World After All. *Current Biology* 24, R959-R962.
- Roesti M, Salzburger W, Berner D (2012b) Uninformative polymorphisms bias genome scans for signatures of selection. *BMC Evolutionary Biology* 12, 94.
- Rundle HD, Nosil P (2005) Ecological speciation. *Ecology Letters* 8, 336-352.
- Sabeti PC, Schaffner SF, Fry B, et al. (2006) Positive Natural Selection in the Human Lineage. *Science* 312, 1614-1620.
- Savolainen O, Lascoux M, Merila J (2013) Ecological genomics of local adaptation. *Nature Reviews Genetics* 14, 807-820.
- Schluter D (2000) *The ecology of adaptive radiation*. Oxford University, Oxford.
- Storz JF (2005) Using genome scans of DNA polymorphism to infer adaptive population divergence. *Molecular Ecology* 14, 671-688.
- Weir BS, Cockerham CC (1984) Estimating F-statistics for the analysis of population-structure. *Evolution* 38, 1358-1370.
- Wright S (1950) Genetical structure of populations. *Nature* 166, 247-249.
- Wu CI (2001) The genic view of the process of speciation. *Journal of Evolutionary Biology* 14, 851-865.

Chapter 1

Genome divergence during evolutionary diversification
as revealed in lake-stream stickleback population pairs

ISSN 0962-1083

VOLUME 21
NUMBER 12
JUNE
2012

MOLECULAR ECOLOGY

FROM THE COVER: Genome divergence during evolutionary diversification as revealed in replicate lake–stream stickleback population pairs. See pp. 2852–2862.



Published by
Wiley-Blackwell

Discover this journal online at
WILEY ONLINE LIBRARY
www.onlinelibrary.com/journal/mec

FROM THE COVER

Genome divergence during evolutionary diversification as revealed in replicate lake–stream stickleback population pairs

MARIUS ROESTI,* ANDREW P. HENDRY,† WALTER SALZBURGER* and DANIEL BERNER*

**Zoological Institute, University of Basel, Vesalgasse 1, CH-4051 Basel, Switzerland, †Department of Biology and Redpath Museum, McGill University, 859 Sherbrooke St W., Montreal, QC, Canada H3A 2K6*

Abstract

Evolutionary diversification is often initiated by adaptive divergence between populations occupying ecologically distinct environments while still exchanging genes. The genetic foundations of this divergence process are largely unknown and are here explored through genome scans in multiple independent lake–stream population pairs of threespine stickleback. We find that across the pairs, overall genomic divergence is associated with the magnitude of divergence in phenotypes known to be under divergent selection. Along this same axis of increasing diversification, genomic divergence becomes increasingly biased towards the centre of chromosomes as opposed to the peripheries. We explain this pattern by within-chromosome variation in the physical extent of hitchhiking, as recombination is greatly reduced in chromosome centres. Correcting for this effect suggests that a great number of genes distributed widely across the genome are involved in the divergence into lake vs. stream habitats. Analyzing additional allopatric population pairs, however, reveals that strong divergence in some genomic regions has been driven by selection unrelated to lake–stream ecology. Our study highlights a major contribution of large-scale variation in recombination rate to generating heterogeneous genomic divergence and indicates that elucidating the genetic basis of adaptive divergence might be more challenging than currently recognized.

Keywords: F_{ST} outlier, *Gasterosteus aculeatus*, gene flow, next generation sequencing, population genomics, RAD, speciation

Received 29 November 2011; revision received 20 January 2012; accepted 24 January 2012

Introduction

Speciation often begins with the adaptive divergence of populations into selectively different ecological environments despite the presence of initially high gene flow (Endler 1977; Schilthuizen 2000; Coyne & Orr 2004; Gavrilets 2004; Via 2009; Sobel *et al.* 2010). The molecular underpinnings of this process remain poorly understood (Wu 2001; Via 2009; Nosil & Schluter 2011). One fundamental unresolved question is how genetic differentiation that builds up between diverging populations is distributed across the genome. Adaptive divergence between populations certainly implies that selection is

strong enough to overcome the homogenizing effect of gene flow at ecologically relevant loci (hereafter 'QTLs') (Wu 2001; Nosil *et al.* 2009; Via 2009). But how many QTLs are involved, how are they arranged across the genome, and how does their divergence influence selectively neutral parts of the genome? Opinions differ widely. At one extreme, some studies argue that, because of hitchhiking, divergence at a few QTLs of major effect can protect large genomic regions from gene flow between selective environments (Turner *et al.* 2005; Via & West 2008; Via 2009). Within these regions, divergence between environments will be elevated relative to the rest of the genome, and additional QTLs can become recruited for further adaptive divergence. At the other extreme, adaptive divergence might involve numerous QTLs of relatively small effect, in which case

Correspondence: Daniel Berner, Fax: +41 (0)61 267 0301; E-mail: daniel.berner@unibas.ch

the hitchhiking of neutral regions along with selected QTLs is predicted to be greatly restricted physically (Barton & Bengtsson 1986; Feder & Nosil 2010). In this case, genetic divergence will either be highly localized or will build up homogeneously throughout the entire genome if reproductive barriers associated with adaptive divergence restrict gene flow effectively enough.

Evaluating the generality of these extreme (as well as intermediate) views on genomic divergence during speciation with gene flow is currently precluded by the scarcity of empirical evidence. The most powerful empirical solutions to this problem are expected to emerge from studies providing high-resolution genome-wide data from multiple replicate population pairs in the initial stages of ecological divergence. These ‘species in waiting’ are particularly informative because the genomic footprints of selection will not yet have been obscured by evolutionary processes acting *after* the completion of reproductive isolation (Coyne & Orr 2004; Via 2009; Nosil & Schluter 2011). Moreover, the incorporation of multiple population pairs differing in their magnitude of divergence allows an explicit examination of how genomic divergence builds up, as opposed to providing only a single temporal snapshot. Our study adopts this approach by combining the power of high-throughput sequencing technology with the availability of an emerging model for studying divergence with gene flow: replicate lake and stream populations of threespine stickleback fish (*Gasterosteus aculeatus*).

Threespine stickleback inhabit contiguous lake and stream habitats in many watersheds that were colonized independently by marine ancestors following the last glacial retreat (Reimchen *et al.* 1985; Lavin & McPhail 1993; Thompson *et al.* 1997; Hendry & Taylor 2004; Berner *et al.* 2009, 2010; Deagle *et al.* 2011). Different lake and stream population pairs typically exhibit similar directions of phenotypic divergence in a number of traits as a response to similar divergent selection (Reimchen *et al.* 1985; Lavin & McPhail 1993; Hendry & Taylor 2004; Berner *et al.* 2009; Kaeuffer *et al.* 2011; Deagle *et al.* 2011). This adaptive divergence likely represents the initial stage of speciation because it frequently coincides with the emergence of at least partial reproductive isolation. (Although it does not necessarily imply that divergence will ever become complete.) In particular, strong shifts in neutral marker allele frequencies occur across lake–stream transitions of just a few hundred metres, even in the absence of physical dispersal barriers (Berner *et al.* 2009). Here, we examine four evolutionarily independent lake and outlet stream stickleback population pairs (‘systems’) from Vancouver Island, Canada. These systems differ in their magnitude of divergence because of differences in the strength of divergent selection, the time for divergence and/or

differences in available genetic variation (Hendry & Taylor 2004; Moore *et al.* 2007; Berner *et al.* 2009; Kaeuffer *et al.* 2011). Importantly, this variation among systems allows us to investigate genomic patterns along a gradient of divergence (Nosil & Schluter 2011). We here present genome scans for all four lake–stream systems based on thousands of markers obtained through Illumina sequencing of restriction site-associated DNA.

Material and methods

Study populations and phenotypic analysis

Our study builds on stickleback sampled from one lake and one outlet stream site in the Boot, Joe’s, Misty, and Robert’s watersheds on Vancouver Island, British Columbia, Canada (sites Boot ‘L’ and ‘S2’, Joe’s ‘L’ and ‘S2’, Misty ‘L’ and ‘S6’, and Robert’s ‘L’ and ‘S2’ in Berner *et al.* 2009). The population pair in each of these systems derives from independent postglacial colonization by marine ancestors (Hendry & Taylor 2004; Berner *et al.* 2009). Absolute barriers to dispersal between lakes and streams are absent in all systems, providing the opportunity for gene flow between the habitats. Details on sampling methods and the populations are provided in Berner *et al.* (2009). This analysis is based on 27 individuals per site (216 in total).

For phenotypic traits, we quantified gill raker number and length, and landmark-based body size and relative body depth, as described in Berner *et al.* (2008, 2011). These traits are known to show strong genetically based divergence between lake and stream populations (Lavin & McPhail 1993; Sharpe *et al.* 2008; Berner *et al.* 2011). We here combined these data into a single multivariate summary metric of within-system phenotypic divergence by mean-scaling each trait and then calculating the Euclidean distance between the lake and the stream sample (univariate patterns are shown in Appendix S1, Supporting information).

Marker generation and quantification of population divergence

To obtain genetic markers, we first prepared libraries of individually barcoded, restriction site-associated DNA (RAD; Baird *et al.* 2008) by largely following the protocol in Hohenlohe *et al.* (2010). Each of the 12 total libraries combined RAD from 18 individuals and was single-end sequenced with 76 cycles in a separate lane on an Illumina genome analyzer Iix. The resulting reads (NCBI short read archive accession number SRP007695) were sorted individually by barcode and then aligned to the reference stickleback genome (Ensembl database version 63.1, assembly Broad S1) by using Novoalign

v2.07.06 (<http://novocraft.com>). We tolerated an equivalent of approximately six high-quality mismatches or gaps and enforced unique alignment, thereby excluding data from repeated elements. Alignments were BAM-converted using Samtools v0.1.11 (Li *et al.* 2009).

For each individual and RAD locus, we then determined the consensus diploid genotype if ten or more replicate reads were available or a haploid consensus genotype if replication was below ten. This threshold was chosen because we identified heterozygote diploids for variable nucleotide positions by a binomial test with insufficient power at low replication. This test involved calculating the binomial likelihood of the observed variant frequency distribution under the null hypothesis of heterozygosity (i.e. assuming a probability of 0.5 for both variants) and accepted heterozygosity if the likelihood was >0.01 . Consensus genotyping was quality aware in that bases with a >0.01 error probability were ignored.

To identify single nucleotide polymorphisms (including a small fraction of microindels, hereafter simply subsumed under 'SNPs'), we pooled the individual consensus genotypes from both habitats within a system at each RAD locus. If a locus was represented by at least 27 consensus genotypes from each habitat (i.e. each individual contributed at least one haplotype on average), we screened every nucleotide position of the locus for variants. Otherwise, the locus was ignored because the quantification of population differentiation was considered unreliable.

Before detected SNPs could be used as genetic markers for analysis, we had to eliminate those lacking the potential to adequately capture the signatures of drift and selection because of a low minor allele frequency. We did so by discarding SNPs with a minor allele frequency of <0.25 (justification and details given in Appendix S2, Supporting information). This filter also effectively eliminated sequencing errors and PCR artefacts from the data but reduced the number of polymorphic RAD loci substantially (e.g. from 12 495 to 4127 in the Boot system). Summary statistics on library size, read coverage, alignment success and marker numbers are provided in Appendix S3 (Supporting information). The remaining (informative) SNPs were then used to calculate F_{ST} based on haplotype diversity (Nei & Tajima 1981, equation 7). For loci harbouring multiple SNPs, we retained for analysis only the one yielding the highest F_{ST} value. However, working with F_{ST} averaged over all SNPs at a given RAD locus, or drawing a single SNP at random, produced similar results supporting identical conclusions in all analyses. Furthermore, using as an alternative divergence metric the chi-square ratio calculated from allele frequencies within a population pair also produced consistent results

throughout, highlighting the robustness of our F_{ST} -based strategy.

Differentiation and recombination rate within chromosomes

Genome-wide F_{ST} patterns suggested a systematic bias of lake–stream divergence towards the chromosome centres (hereafter called 'chromosome centre-biased divergence', CCBD; see Results). To formally quantify this observation, we divided each chromosome physically into its 'centre' (inner 50% of a chromosome's sequence) and its 'peripheries' (outer 25% on each side). We then subtracted mean F_{ST} of all markers in the periphery from mean F_{ST} of the markers in the centre and calculated the mean and 95% confidence interval for this CCBD metric within each lake–stream system by using all chromosomes as data points ($N = 21$).

To explore whether CCBD was associated with recombination rate, we extracted information on genetic (linkage) distance (in cM) and physical distance (in mb) for the SNPs and microsatellite markers underlying the stickleback linkage maps presented in Albert *et al.* (2008) and Greenwood *et al.* (2011). The ratio of genetic by physical distance for neighbouring markers then provided an estimate of the average recombination rate for that marker interval (Appendix S4, Supporting information). In addition, we used information on the physical location of the centromere on each chromosome (Urton *et al.* 2011) to evaluate whether heterogeneity in divergence and recombination rate along chromosomes was related to centromere position.

Sliding window analysis screening for outlier regions

The magnitude of population divergence at a given locus proved dependent on chromosome position at a large physical scale (CCBD). Screening for localized regions of high divergence (F_{ST} 'outliers'), potentially indicating hitchhiking along with QTLs under divergent selection (Storz 2005; Nielsen 2005), thus first required an adjustment of F_{ST} values to account for CCBD (see Discussion). To do so, we subjected system- and chromosome-specific F_{ST} data to locally weighted scatterplot smoothing ('LOESS', a nonparametric regression) with chromosome position as predictor. (The polynomial degree was zero in all analyses; hence, LOESS produced a moving average). We used a relatively high bandwidth (0.3) to capture only the coarse heterogeneity in divergence within a chromosome. We then calculated 'residual divergence' at each marker as the difference between the raw and the fitted F_{ST} values. Because CCBD increased with overall divergence (see below),

this procedure had a large effect in the Boot system but a relatively minor effect in the other systems.

To explore the number and physical arrangement of outlier regions, *residual* divergence within each system was subjected to sliding window analysis using LOESS with a narrow bandwidth (0.03) facilitating visualization while adequately conserving small-scale divergence heterogeneity along each chromosome. We excluded the Misty system from this analysis because we suspected a low signal to noise ratio in this barely differentiated lake–stream pair. Outlier significance thresholds were determined empirically based on a resampling strategy (Appendix S5, Supporting information). In addition to the ‘parapatric’ lake–stream comparisons within each system, we also performed ‘allopatric’ comparisons between populations of the same habitat type (i.e. lake–lake and stream–stream population pairings). Parapatric vs. allopatric comparisons then allowed us to compare patterns of genomic divergence across different ecological settings.

All analyses except for sequence alignment were performed in R (R Development Core Team 2010), making use of the R-Bioconductor packages ShortRead (Morgan *et al.* 2009), Rsamtools, and Biostrings.

Results

We found dramatic differences among systems in the magnitude of overall baseline genomic divergence (Fig. 1), and this paralleled the magnitude of differences among systems in phenotypic divergence (see red and blue bars in Fig. 2). In particular, approximately 0.6% of the markers in the most divergent system (Boot) reached fixation of alternative variants between the habitats. By contrast, no locus reached appreciable divergence in the Misty system. Furthermore, heterogeneity in divergence along the genome increased with increasing baseline divergence (Fig. 1).

As noted earlier, a striking pattern towards higher F_{ST} values in the chromosome centres than in the chromosome peripheries was evident, particularly in the Boot system. A metric based on the difference in mean F_{ST} between markers from the centre and from the peripheries of each chromosome confirmed this pattern (Fig. 2), which we call ‘chromosome centre-biased divergence’ (CCBD). CCBD averaged across chromosomes within systems was related to the overall magnitude of phenotypic and baseline genetic lake–stream divergence in those systems: that is, CCBD was absent in the undifferentiated Misty system but was very

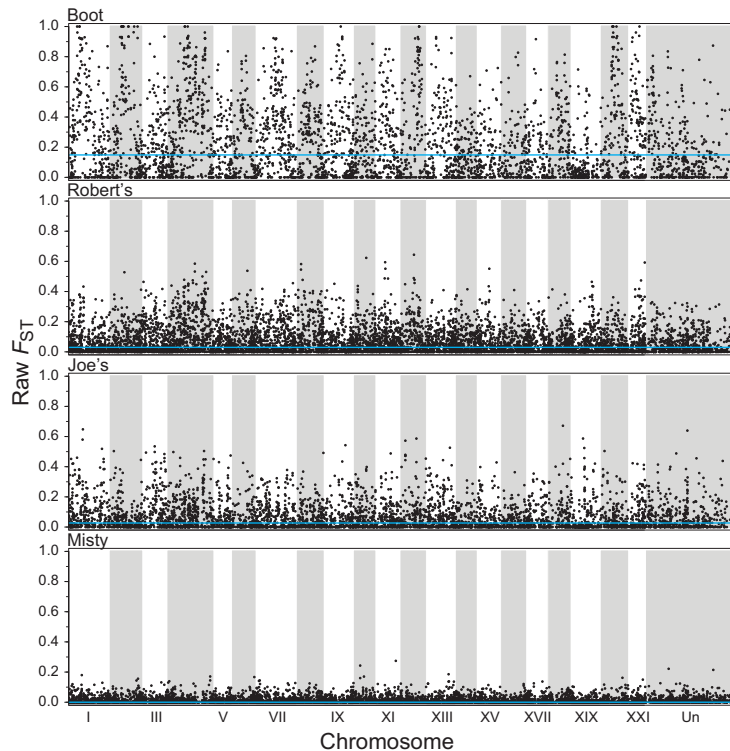


Fig. 1 Genome-wide divergence in four independent population pairs (systems) of lake and stream stickleback. The dots show F_{ST} values for each marker on each chromosome; the chromosomes are separated by white and grey background shading. [‘Un’ is the artificial chromosome consisting of concatenated unanchored scaffolds. Also, chromosome XIX was corrected for misassembly (Ross & Peichel 2008) in all analyses.] Total marker coverage per system ranges between 4127 and 8417 (Appendix S3, Supporting information). The blue horizontal line represents baseline divergence defined as genome-wide median F_{ST} (Misty: 0; Joe’s: 0.027; Robert’s: 0.030; Boot: 0.149). Moving from the bottom (Misty) to the top (Boot), note increasing magnitudes of baseline divergence, and increasing heterogeneity in divergence across the genome.

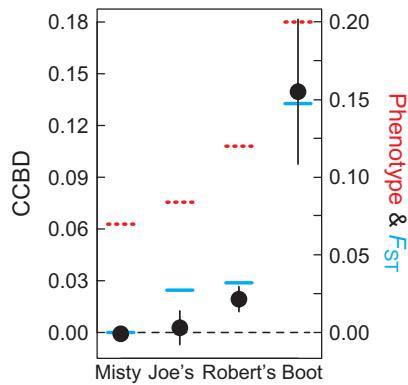


Fig. 2 The emergence of chromosome centre-biased divergence (CCBD). CCBD is expressed as the difference between the chromosome centre (inner 50% of the sequence) and the chromosome peripheries (outer 50%) in the magnitude of differentiation (F_{ST}) between the lake and stream habitat within each system. Dots and error bars are means and 95% confidence intervals across the 21 chromosomes. Positive values indicate relatively greater divergence in the centre of the chromosomes as opposed to their peripheries. CCBD emerges when divergence becomes substantial, as quantified by phenotypic divergence and genome-wide median F_{ST} (dashed red and solid blue horizontal bars, both referring to the right axis). Phenotypic divergence integrates four ecologically important and genetically based morphological traits (Appendix S1, Supporting information).

strong in the Boot system exhibiting greatest progress in divergence (the top row in Fig. 3 shows a fine-scale illustration of CCBD for two exemplary chromosomes in the Boot system; patterns on all chromosomes for that system are presented in supporting online Appendix S6a, Supporting information).

All chromosomes for which enough data were available exhibited a valley of reduced recombination around their centre (Fig. 3, middle row; Appendix S4, Supporting information). The variation in recombination rate was often dramatic, with a 10 fold or higher reduction in the centre of some chromosomes relative to their peripheries (see e.g. chromosome VII, Fig. 3). Both CCBD and physical variation in recombination rate were unrelated to the position of the centromere (Fig. 3; Appendix S4, Supporting information).

After adjusting raw F_{ST} values for CCBD (residual F_{ST} ; Fig. 3, bottom), our sliding window analyses found outlier regions in relatively high numbers throughout the genome in all three systems (Fig. 4; Misty excluded owing to the overall lack of differentiation). A qualitative comparison indicated that outlier regions were relatively inconsistent across the systems. For instance, we found no peak exceeding the $P < 0.01$ threshold in all three systems. Significant outliers were

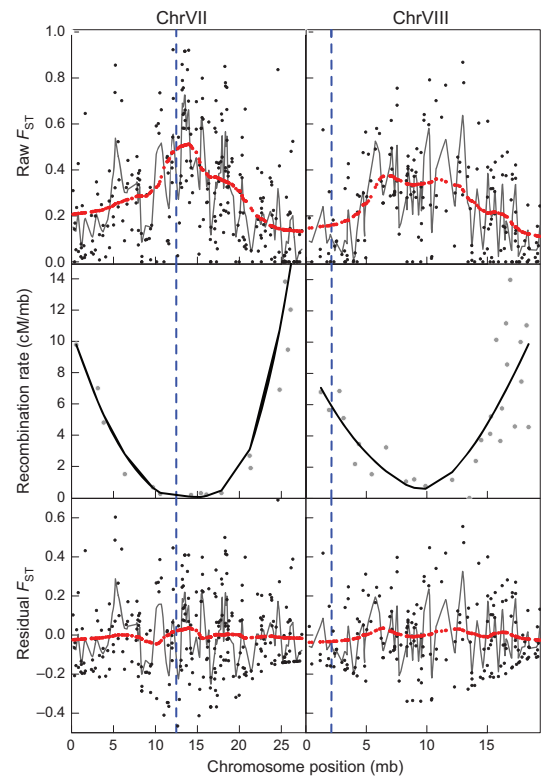


Fig. 3 *Top panels.* Lake–stream divergence in the Boot system along chromosomes seven and eight. Black dots give the raw F_{ST} value at each marker, the grey line connects F_{ST} values predicted by a fine-scale smoother (LOESS, bandwidth = 0.03), and red dots represent F_{ST} values predicted by a coarse smoother (bandwidth = 0.3). These data illustrate that population divergence is greater in the chromosome centres than in their peripheries (CCBD). Divergence profiles for all other chromosomes in the Boot system are presented in Appendix S6a (Supporting information). *Middle panels.* Recombination rates for marker intervals along the same two chromosomes show that recombination is dramatically lower in the chromosome centres relative to the peripheries. Note that heterogeneity along the chromosomes in both F_{ST} and recombination rate is independent from the position of the centromere, indicated by the dashed blue vertical line (chromosome seven is metacentric, whereas chromosome eight is telocentric; Urton *et al.* 2011). Recombination rates and centromere positions for the other chromosomes are presented in Appendix S4 (Supporting information). *Bottom panels.* Separating locus-specific signatures of selection from CCBD. The plotting conventions are as in the top row, except that the underlying data points are residual F_{ST} obtained by subtracting the values predicted by the coarse smoother (red dots) in the top panels from the raw F_{ST} values (black dots) in the top panels.

also observed in allopatric population comparisons, with an exemplar allopatric comparison shown in Fig. 4 (bottom). Interestingly, some outlier regions

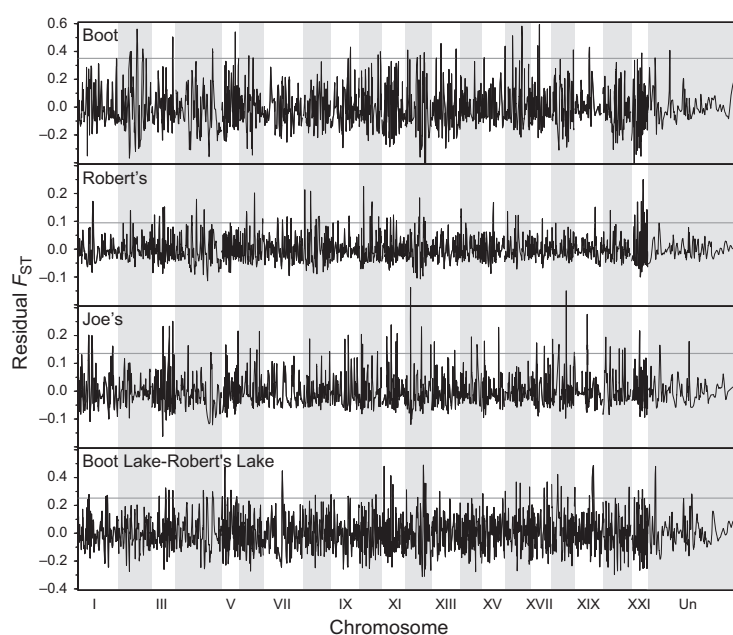


Fig. 4 Sliding window analyses visualizing genome-wide divergence. The top three panels show lake–stream comparisons within the three divergent systems (the Misty system was excluded, because divergence was minimal). The bottom panel shows an exemplary allopatric population comparison (Boot Lake–Robert’s Lake; 8735 markers, median $F_{ST} = 0.266$). The divergence profiles are based on residual (CCBD-corrected) F_{ST} (see Fig. 3). Grey horizontal lines indicate $P < 0.01$ significance thresholds for outlier regions determined by a resampling approach (Appendix S5, Supporting information). Note that the relatively weakly divergent Robert’s and Joe’s systems are plotted on a two-fold finer scale than the Boot and the allopatric comparisons.

suggested by the parapatric lake–stream comparisons also emerged in allopatric comparisons. A particularly clear case involves the two high-divergence peaks flanking the low-divergence *Ectodysplasin* (*Eda*) locus at a distance of 1–2 mb (Fig. 5).

Discussion

We used stickleback population pairs from lake and stream habitats in multiple independent watersheds to characterize how genomes diverge when populations diversify in the face of gene flow. One major finding is that striking differences are evident among systems in the overall magnitude of lake–stream genomic divergence (Fig. 1) and that these differences match those previously documented for phenotypes and microsatellites (Berner *et al.* 2009). In particular, while baseline divergence is substantial and a number of markers have reached fixation for alternative variants in the Boot system, divergence is weaker in the Robert’s and Joe’s systems, and negligible in the Misty system. Given that gene flow is known to be very high from Misty Lake into the Misty outlet stream, despite evidence for strong divergent selection (Hendry *et al.* 2002; Moore *et al.* 2007; Berner *et al.* 2009), our analysis here provides a robust demonstration of genome-wide constraints on adaptive divergence as a result of homogenizing gene flow. That is, gene flow in the Misty system overwhelms divergence even for the loci likely subject to the strongest divergent selection.

Chromosome centre-biased divergence (CCBD)

Another major finding is that increasing phenotypic and genetic divergence leads to relatively stronger divergence in chromosome centres than towards their peripheries (CCBD), which contributes to increasing heterogeneity (or variance) in divergence across the genome. The most straightforward explanation for CCBD is the coincidence of adaptive divergence at multiple QTLs and reduced recombination rate in the chromosome centres. The reason is that, for a given magnitude of divergence at a QTL, associated hitchhiking will extend deeper into the neutral neighbourhood if the QTL is located in a genomic region where the recombination rate is relatively reduced (Barton & Bengtsson 1986; Kaplan *et al.* 1989; Charlesworth *et al.* 1997; Feder & Nosil 2010). Moreover, CCBD will be particularly pronounced if a chromosome harbours multiple QTLs under divergent selection, because the hitchhiking effect of the QTLs will tend to cumulate more strongly in the centre than in the periphery. Our analysis of recombination rate based on stickleback linkage maps is consistent with this hypothesized mechanism: stickleback chromosomes consistently display reduced recombination in their centres relative to their peripheries (see also Hohenlohe *et al.* 2012). Similar within-chromosome variation in recombination rate has recently been reported from several genetic model organisms [*C. elegans*: Rockman & Kruglyak (2009); zebrafish: Bradley *et al.* (2011); mice, rats, humans: Jensen-Seaman *et al.*

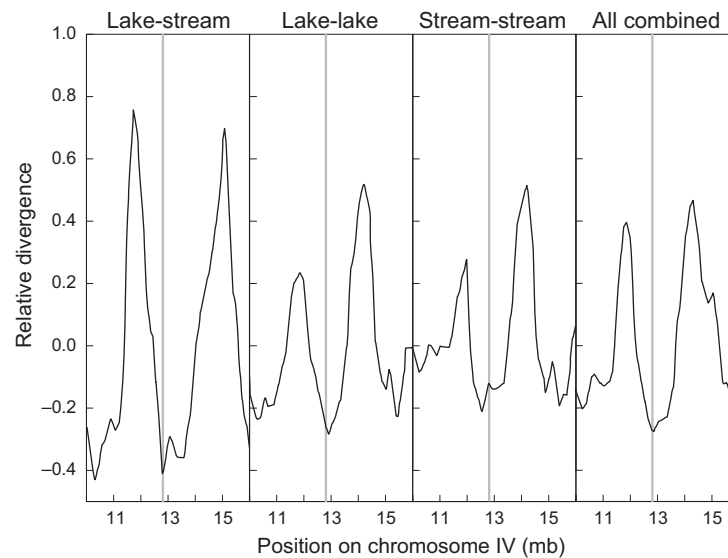


Fig. 5 Divergence profiles along a 6 mb segment of chromosome IV centred at the *Ectodysplasin* (*Eda*) locus (grey vertical line), which is the key genetic factor in lateral plate reduction (Colosimo *et al.* 2005). Profiles are shown for parapatric (lake-stream) and allopatric (lake-lake, stream-stream) population comparisons, and for an analysis combining all these data. The parapatric class averages divergence data from all four lake-stream systems, while each allopatric class averages data from all six possible population pairings. The combined analysis thus integrates data from 16 total population comparisons, yielding a total of 261 markers and an average inter-marker spacing of 23 kb within the displayed chromosome segment. Data averaging was made possible by expressing divergence in each population comparison in units of baseline divergence ('relative divergence'; obtained by scaling residual divergence by genome-wide mean F_{ST}). Note that *Eda* consistently displays below-baseline divergence but is flanked by a high-divergence peak on each side.

(2004); Borodin *et al.* (2008); Chowdhury *et al.* (2009)]. The emergence of CCBD during adaptive divergence might thus be a common phenomenon.

While the reason for within-chromosome variation in recombination rate remains unclear—it appears unrelated to the position of the centromere—the phenomenon has at least two important implications. First, the associated CCBD challenges the conceptual dichotomy between divergence beginning with the emergence of a few large and isolated differentiated regions associated with large-effect QTLs (Via & West 2008; Via 2009), vs. more homogeneous genome-wide divergence associated with numerous QTLs of minor effect (Feder & Nosil 2010). That is, given reduced recombination in chromosome centres, even minor-effect QTLs might drive strong marker divergence over large genomic regions when they happen to co-localize in chromosome centres, whereas large-effect QTLs might not generate much hitchhiking when located in the highly recombining peripheries. Our study thus highlights a key role of variable recombination rate in generating heterogeneous genomic divergence during evolutionary diversification and indicates that the prevailing focus on pericentric regions and inversions (Butlin 2005; Kirkpatrick & Barton 2006; Hoffmann & Rieseberg 2008; Feder & Nosil

2009; Noor & Bennett 2009) misses important variation in recombination rate at a much larger physical scale.

A second implication of within-chromosome variation in recombination rate and CCBD is methodological. Because hitchhiking is expected to be more extensive in chromosome centres, the probability of a particular marker detecting the signature of a locus under selection is relatively higher in the chromosome centres. In addition, genomic regions under divergent selection in nonmodel organisms are often identified by anonymous genome scans that do not map markers to a reference genome or a linkage map (e.g. Beaumont & Balding 2004; Foll & Gaggiotti 2008; Excoffier *et al.* 2009). These approaches assume that locus-specific F_{ST} values can be evaluated against a *genome-wide* baseline. CCBD undermines this assumption and hence leads to a systematic bias towards identifying outliers at markers located near chromosome centres. That is, anonymous genome scans cannot separate localized signatures of hitchhiking associated with *specific* selected QTLs from diffuse, large-scale heterogeneity in divergence along chromosomes driven by *multiple* selected QTLs and large-scale reduced recombination. Our strategy to address this problem was to express divergence at each marker as the deviation of the raw F_{ST} value from the F_{ST} value

predicted by a coarse smoothing function capturing CCBD (yielding ‘residual divergence’; Fig. 3). We do not claim that this *ad hoc* empirical standardization is optimal. Until more sophisticated methods are developed, however, localized signatures of selection in systems exhibiting CCBD are certainly better inferred based on residual F_{ST} than on raw F_{ST} .

Outlier analysis

Our sliding window analyses based on residual (CCBD-adjusted) divergence suggested the presence of dozens of outliers spread throughout the genome within each system. Many loci thus appear to contribute to adaptive divergence, as has also recently been inferred for *Anopheles* mosquitoes (Lawniczak *et al.* 2010), and marine vs. freshwater stickleback (Hohenlohe *et al.* 2010). This finding contradicts the idea that during the early stages of speciation, divergence builds up in only a few genomic hotspots associated with major QTLs (Via & West 2008; Via 2009). Our inference of numerous selected QTLs is also consistent with the observation of CCBD; if only a few loci were targeted by selection, strong and consistent CCBD would not be expected to emerge.

We also found that divergence profiles were rather inconsistent among our systems, making it difficult to identify genetic regions of *general* importance to lake–stream stickleback divergence. A similar conclusion was reached in a recent lower-resolution genome scan using several lake–stream stickleback populations from another region of British Columbia (Deagle *et al.* 2011). Possible explanations include differences in the nature of divergent lake–stream selection among the systems (Berner *et al.* 2008, 2009; Kaeuffer *et al.* 2011), or the possibility that responses to similar divergent selection involve different QTLs in the different systems (Arendt & Reznick 2008; Kaeuffer *et al.* 2011). The latter would not be surprising, as many traits involved in adaptive divergence between lakes and streams are likely polygenic (Peichel *et al.* 2001; Albert *et al.* 2008; Greenwood *et al.* 2011).

It is also possible, however, that the inconsistency in outliers among systems reflects a fundamental limitation of genome scans. Drawing on theory (Slatkin & Wiehe 1998; Barton 2000; Bierne 2010), we predict that the recurrent fixation of an *unconditionally* favourable QTL allele from the standing genetic variation will generate peaks of high population divergence in neutral regions flanking the QTL on both sides, while the QTL itself will remain undifferentiated. The reason is that different copies of the favourable allele will share their immediate neutral neighbourhood, while potentially being associated with different neutral variants further away from the QTL. The pattern we found at the *Eda* locus across parapatric and allopatric population comparisons (Fig. 5) is consis-

tent with this scenario. *Eda* is the key genetic factor underlying adaptive lateral plate reduction following freshwater colonization by stickleback, and all of our (low-plated) populations are likely fixed for the same derived *Eda* allele available in the ancestral standing variation of the colonizing marine fish (Colosimo *et al.* 2005). The twin peaks flanking *Eda* therefore reflect hitchhiking with a single unconditionally favourable allele (i.e. an allele favoured in *both* lakes and streams) rather than two separate signatures of divergent lake–stream selection. Interestingly, Deagle *et al.* (2011) inferred a locus presumably influenced by divergent lake–stream selection almost exactly at the tip of the left peak flanking *Eda* in our analysis (at 12 mb; see the first marker in their Table 2). Similarly, Jones *et al.* (2012) interpreted two outlier regions flanking *Eda* (at 11.4 mb and 15.7 mb; see their Fig. 3) as indicating loci involved in the divergence of sympatric benthic–limnetic stickleback. In the light of our findings, these interpretations need to be revised.

Overall, the conclusion that lake–stream divergence involves numerous QTLs is probably robust. However, the above-mentioned considerations (Slatkin & Wiehe 1998; Bierne 2010) and results highlight that regions of high divergence identified in (replicate) genome scans are not necessarily related to *divergent* selection mediated by the causal factor of interest (here lake–stream ecology). Allele frequency shifts at QTLs driven by *any* type of selection within a local population can generate outliers in linked markers between populations (Charlesworth *et al.* 1997; Charlesworth 1998).

Conclusions

Our genome scan comparisons of multiple lake–stream stickleback population pairs have shown that increasing phenotypic divergence coincides with increasing overall genomic divergence, and with increasing large-scale heterogeneity in divergence across the genome. Heterogeneous divergence is strongly driven by within-chromosome variation in recombination rate, a phenomenon that might be common and hence requires conceptual integration in speciation genetics. Large-scale heterogeneous divergence also represents an unappreciated methodological challenge to genome scans in search for selected loci. Our study further suggests that lake–stream divergence involves shifts at numerous QTLs throughout the genome but also cautions that inferring the selective context underlying regions of high divergence is less straightforward than generally recognized.

Acknowledgements

We thank A.-C. Grandchamp, J.-S. Moore and K. Hudson for aiding field work; P. Etter and W. Cresko for sharing their

expertise in RAD library preparation; B. Aeschbach and N. Boileau for facilitating wet laboratory work; I. Nissen for Illumina sequencing at the Quantitative Genomics Facility, D-BSSE, ETH Zürich; L. Zimmermann for IT support; C. Hercus for modifying Novoalign; M. Morgan and H. Pagès for coding suggestions; D. Ebert, C. Peichel, H. Hoekstra, H. Gante and three referees for valuable comments and suggestions on data analysis; J. Urton for details on centromere positions; M. Hansen and T. Vines for efficient manuscript handling. APH was funded by the Natural Sciences and Engineering Research Council (NSERC) of Canada, WS by the European Research Council (ERC, Starting Grant 'INTERGENADAPT'), the Swiss National Science Foundation and the University of Basel, and DB by the Swiss National Science Foundation (PBB5A-111216 and Ambizione PZ00P3_126391/1).

References

- Albert AYK, Sawaya S, Vines TH *et al.* (2008) The genetics of adaptive shape shift in stickleback: pleiotropy and effect size. *Evolution*, **62**, 76–85.
- 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.
- Baird NA, Etter PD, Atwood TS *et al.* (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE*, **3**, e3376.
- Barton NH (2000) Genetic hitchhiking. *Philosophical Transactions of the Royal Society of London, Series B, Biological Sciences*, **355**, 1553–1562.
- Barton N, Bengtsson BO (1986) The barrier to genetic exchange between hybridizing populations. *Heredity*, **57**, 357–376.
- Beaumont MA, Balding DJ (2004) Identifying adaptive genetic divergence among populations from genome scans. *Molecular Ecology*, **13**, 969–980.
- Berner D, Adams DC, Grandchamp A-C, Hendry AP (2008) Natural selection drives patterns of lake-stream divergence in stickleback foraging morphology. *Journal of Evolutionary Biology*, **21**, 1653–1665.
- Berner D, Grandchamp A-C, Hendry AP (2009) Variable progress toward ecological speciation in parapatry: stickleback across eight lake-stream transitions. *Evolution*, **63**, 1740–1753.
- Berner D, Roesti M, Hendry AP, Salzburger W (2010) Constraints on speciation suggested by comparing lake-stream stickleback divergence across two continents. *Molecular Ecology*, **19**, 4963–4978.
- Berner D, Kaeuffer R, Grandchamp A-C *et al.* (2011) Quantitative genetic inheritance of morphological divergence in a lake-stream stickleback ecotype pair: implications for reproductive isolation. *Journal of Evolutionary Biology*, **24**, 1975–1983.
- Bierne N (2010) The distinctive footprints of local hitchhiking in a varied environment and global hitchhiking in a subdivided population. *Evolution*, **64**, 3254–3272.
- Borodin PM, Karamysheva TV, Belonogova NM *et al.* (2008) Recombination map of the common shrew, *Sorex araneus* (eulipotyphla, mammalia). *Genetics*, **178**, 621–632.
- Bradley KM, Breyer JP, Melville DB *et al.* (2011) An SNP-based linkage map for zebrafish reveals sex determination loci. *G3: Genes, Genomes, Genetics*, **1**, 3–9.
- Butlin RK (2005) Recombination and speciation. *Molecular Ecology*, **14**, 2621–2635.
- Charlesworth B (1998) Measures of divergence between populations and the effect of forces that reduce variability. *Molecular Biology and Evolution*, **15**, 538–543.
- Charlesworth B, Nordborg M, Charlesworth D (1997) The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations. *Genetical Research*, **70**, 155–174.
- Chowdhury R, Bois PRJ, Feingold E, Sherman SL, Cheung VG (2009) Genetic analysis of variation in human meiotic recombination. *PLoS Genetics*, **5**, e1000648.
- Colosimo PF, Hosemann KE, Balabhadra S *et al.* (2005) Widespread parallel evolution in sticklebacks by repeated fixation of ectodysplasin alleles. *Science*, **307**, 1928–1933.
- Coyne JA, Orr HA (2004) *Speciation*. Sinauer Associates, Sunderland, Massachusetts.
- Deagle BE, Jones FC, Chan YF *et al.* (2011) Population genomics of parallel phenotypic evolution in stickleback across stream-lake ecological transitions. *Proceedings of the Royal Society of London Series B Biological Sciences*, in press.
- Endler JA (1977) *Geographic Variation, Speciation, and Clines*. Princeton University, Princeton, New Jersey.
- Excoffier L, Hofer T, Foll M (2009) Detecting loci under selection in a hierarchically structured population. *Heredity*, **103**, 285–298.
- Feder JL, Nosil P (2009) Chromosomal inversions and species differences: when are genes affecting adaptive divergence and reproductive isolation expected to reside within inversions? *Evolution*, **63**, 3061–3075.
- Feder JL, Nosil P (2010) The efficacy of divergence hitchhiking in generating genomic islands during ecological speciation. *Evolution*, **64**, 1729–1747.
- Foll M, Gaggiotti O (2008) A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. *Genetics*, **180**, 977–993.
- Gavrilets S (2004) *Fitness Landscapes and the Origin of Species*. Princeton University, Princeton, New Jersey.
- Greenwood AK, Jones FC, Chan YF *et al.* (2011) The genetic basis of divergent pigment patterns in juvenile threespine sticklebacks. *Heredity*, **107**, 155–166.
- Hendry AP, Taylor EB (2004) How much of the variation in adaptive divergence can be explained by gene flow? An evaluation using lake-stream stickleback pairs. *Evolution*, **58**, 2319–2331.
- Hendry AP, Taylor EB, McPhail JD (2002) Adaptive divergence and the balance between selection and gene flow: lake and stream stickleback in the Misty system. *Evolution*, **56**, 1199–1216.
- Hoffmann AA, Rieseberg LH (2008) Revisiting the impact of inversions in evolution: from population genetic markers to drivers of adaptive shifts and speciation? *Annual Review of Ecology and Systematics*, **39**, 21–42.
- Hohenlohe PA, Bassham S, Etter PD *et al.* (2010) Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLoS Genetics*, **6**, e1000862.
- Hohenlohe PA, Bassham S, Currey M, Cresko WA (2012) Extensive linkage disequilibrium and parallel adaptive divergence across threespine stickleback genomes. *Philosophical Transactions of the Royal Society B, Biological Sciences*, **367**, 395–408.

- Jensen-Seaman MI, Furey TS, Payseur BA *et al.* (2004) Comparative recombination rates in the rat, mouse, and human genomes. *Genome Research*, **14**, 528–538.
- Jones FC, Chan YF, Schmutz J *et al.* (2012) A genome-wide SNP genotyping array reveals patterns of global and repeated Species-pair divergence in sticklebacks. *Current Biology*, **22**, 83–90.
- Kaeuffer R, Peichel C, Bolnick DI, Hendry AP (2011) Parallel and nonparallel aspects of ecological, phenotypic, and genetic divergence across replicate population pairs of lake and stream stickleback. *Evolution*, **66**, 402–418.
- Kaplan NL, Hudson RR, Langley CH (1989) The hitchhiking effect revisited. *Genetics*, **123**, 887–899.
- Kirkpatrick M, Barton N (2006) Chromosome inversions, local adaptation and speciation. *Genetics*, **173**, 419–434.
- Lavin PA, McPhail JD (1993) Parapatric lake and stream sticklebacks on northern Vancouver Island: disjunct distribution or parallel evolution? *Canadian Journal of Zoology*, **71**, 11–17.
- Lawniczak MKN, Emrich SJ, Holloway AK *et al.* (2010) Widespread divergence between incipient *Anopheles gambiae* species revealed by whole genome sequences. *Science*, **330**, 512–514.
- Li H, Handsaker B, Wysoker A *et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
- Moore JS, Gow JL, Taylor EB, Hendry AP (2007) Quantifying the constraining influence of gene flow on adaptive divergence in the lake-stream threespine stickleback system. *Evolution*, **61**, 2015–2026.
- Morgan M, Anders S, Lawrence M *et al.* (2009) ShortRead: a bioconductor package for input, quality assessment and exploration of high-throughput sequence data. *Bioinformatics*, **25**, 2607–2608.
- Nei M, Tajima F (1981) DNA polymorphism detectable by restriction endonucleases. *Genetics*, **97**, 145–163.
- Nielsen R (2005) Molecular signatures of natural selection. *Annual Review of Genetics*, **39**, 197–218.
- Noor MAF, Bennett SM (2009) Islands of speciation or mirages in the desert? Examining the role of restricted recombination in maintaining species *Heredity*, **103**, 439–444.
- Nosil P, Schluter D (2011) The genes underlying the process of speciation. *Trends in Ecology and Evolution*, **26**, 160–167.
- Nosil P, Funk DJ, Ortiz-Barrientos D (2009) Divergent selection and heterogeneous genomic divergence. *Molecular Ecology*, **18**, 375–402.
- Peichel CL, Nereng KS, Oghi KA *et al.* (2001) The genetic architecture of divergence between threespine stickleback species. *Nature*, **414**, 901–905.
- R Development Core Team (2010) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Reimchen TE, Stinson EM, Nelson JS (1985) Multivariate differentiation of parapatric and allopatric populations of threespine stickleback in the Sangan River watershed, Queen Charlotte Islands. *Canadian Journal of Zoology*, **63**, 2944–2951.
- Rockman MV, Kruglyak L (2009) Recombinational landscape and population genomics of *Caenorhabditis elegans*. *PLoS Genetics*, **5**, e1000419.
- Ross JA, Peichel CL (2008) Molecular cytogenetic evidence of rearrangements on the Y chromosome of the threespine stickleback fish. *Genetics*, **179**, 2173–2182.
- Schilthuizen M (2000) Dualism and conflicts in understanding speciation. *BioEssays*, **22**, 1134–1141.
- Sharpe DMT, Räsänen K, Berner D, Hendry AP (2008) Genetic and environmental contributions to the morphology of lake and stream stickleback: implications for gene flow and reproductive isolation. *Evolutionary Ecology Research*, **10**, 849–866.
- Slatkin M, Wiehe T (1998) Genetic hitch-hiking in a subdivided population. *Genetical Research*, **71**, 155–160.
- Sobel JM, Chen GF, Watt LR, Schemske DW (2010) The biology of speciation. *Evolution*, **64**, 295–315.
- Storz JF (2005) Using genome scans of DNA polymorphism to infer adaptive population divergence. *Molecular Ecology*, **14**, 671–688.
- Thompson CE, Taylor EB, McPhail JD (1997) Parallel evolution of lake-stream pairs of threespine sticklebacks (*Gasterosteus*) inferred from mitochondrial DNA variation. *Evolution*, **51**, 1955–1965.
- Turner TL, Hahn MW, Nuzhdin SV (2005) Genomic islands of speciation in *Anopheles gambiae*. *PLoS Biology*, **3**, e285.
- Urton JR, McCann SM, Peichel CL (2011) Karyotype differentiation between two stickleback species (*Gasterosteidae*). *Cytogenetic and Genome Research*, **135**, 150–159.
- Via S (2009) Natural selection in action during speciation. *Proceedings of the National Academy of Sciences of the United States of America*, **106**, 9939–9946.
- Via S, West J (2008) The genetic mosaic suggests a new role for hitchhiking in ecological speciation. *Molecular Ecology*, **17**, 4334–4345.
- Wu CI (2001) Genes and speciation. *Journal of Evolutionary Biology*, **14**, 889–891.

M.R. is a PhD student in the Salzburger laboratory and has general interest in the processes underlying biological diversification. He is currently investigating diversification in lake-stream stickleback from population genomic and ecological angles. A.H. investigates factors that influence the evolution of biological diversity, including natural selection, gene flow, adaptation and reproductive isolation. He conducts research in a number of study systems, including Darwin's finches (Galápagos Islands), guppies (Trinidad and Tobago) and stickleback (British Columbia). W.S. is Assistant Professor at the Zoological Institute of the University of Basel. The research of his team focuses on the genetic basis of adaptation, evolutionary innovation and animal diversification. The laboratory's homepage at <http://www.evolution.unibas.ch/salzburger/> provides further details on the group's (research) activities. D.B. is interested in adaptation and uses lake and stream stickleback populations for his empirical research.

Data accessibility

DNA sequences: NCBI short read archive accession number SRP007695. Morphological data for Berner *et al.* (2010): Dryad digital repository (doi:10.5061/dryad.1960).

Supporting information

Appendix S1 Univariate divergence in four morphological traits within each lake–stream system.

Appendix S2 Strategy adopted to eliminate uninformative polymorphisms from the marker data sets.

Appendix S3 Summary statistics on library size, read coverage, alignment success and marker numbers.

Appendix S4 Recombination rate along stickleback chromosomes.

Appendix S5 Resampling approach used to determine significance thresholds for F_{ST} outlier regions.

Appendix S6 Genome-wide sliding window profile of lake-stream divergence within the Boot system (both raw F_{ST} and residual F_{ST}).

Please note: Wiley-Blackwell is not responsible for the content or functionality of any supporting information supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

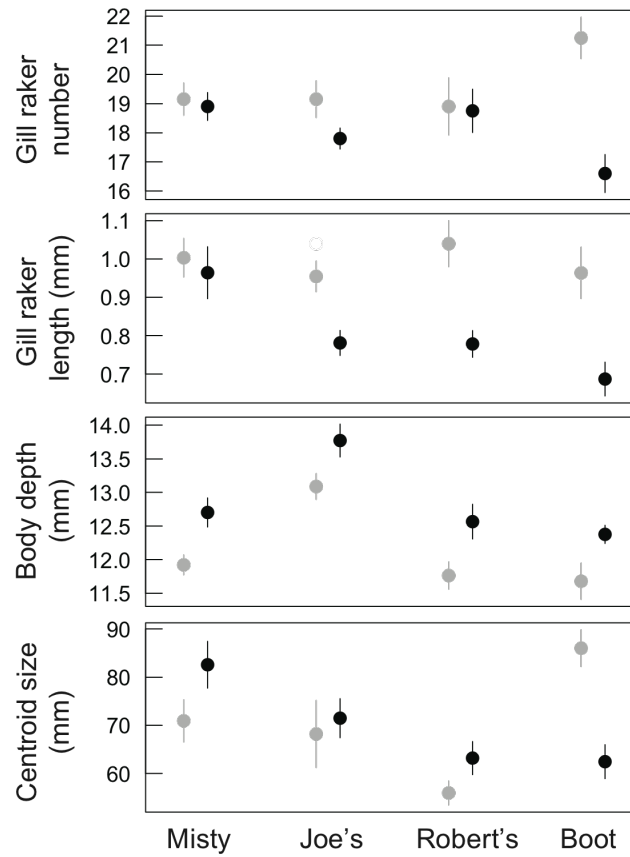
Chapter 1

Genome divergence during evolutionary diversification
as revealed in lake-stream stickleback population pairs

Supplementary material

Supporting Appendix S1 to Roesti et al. 2012

Univariate divergence in morphology between lake and stream stickleback (gray and black symbols) from the four lake-stream systems. Error bars are 95% confidence intervals. After mean-scaling these phenotypes, Euclidean distance was calculated to obtain the multivariate divergence metric used in Fig. 2.



Supporting Appendix S2 to Roesti et al. 2012

Polymorphisms with a low minor allele frequency are constrained to exhibit low F_{st} values *irrespective* of the evolutionary processes operating in their genomic neighborhood, and therefore introduce bias into genome scan studies. To find an adequate strategy to exclude such SNPs from our data sets, we examined how the correlation in the magnitude of lake-stream divergence (F_{st} ; Nei & Tajima 1981) between SNPs on 'sister' RAD loci (i.e., SNPs associated with the *same* genomic restriction site and hence separated physically by between 9 and 135 bases only) was influenced by imposing increasingly severe minor allele frequency filters. The rationale was that due to the extreme physical proximity of sister RAD loci, drift and hitchhiking should influence polymorphisms on these loci similarly, leading to similar F_{st} values. A poor correlation in F_{st} between SNPs on sister loci would therefore indicate a relatively high proportion of uninformative SNPs that estimate divergence imprecisely. This was examined in the Boot system, displaying the widest range of F_{st} values.

Indeed, we found a poor correlation ($r = 0.386$, $N = 2103$) between F_{st} values estimated from SNPs on sister loci when accepting SNPs with *any* minor allele frequency (including singletons) in the analysis (Fig. A1a). Successive elimination of SNPs through increasingly restrictive minor allele frequency thresholds improved the correlation dramatically (Fig. A1b). We finally settled on a frequency filtering threshold of 0.25 for the minor allele. That is, the minor allele had to account for at least 25% of the total base count (minor and major allele combined) at a SNP (lake and stream sample pooled). This rule, leading to a high F_{st} correlation between sister loci ($r = 0.842$, $N = 356$) (Fig. A1c), represented a reasonable compromise between high marker coverage and reliable divergence estimation by the markers.

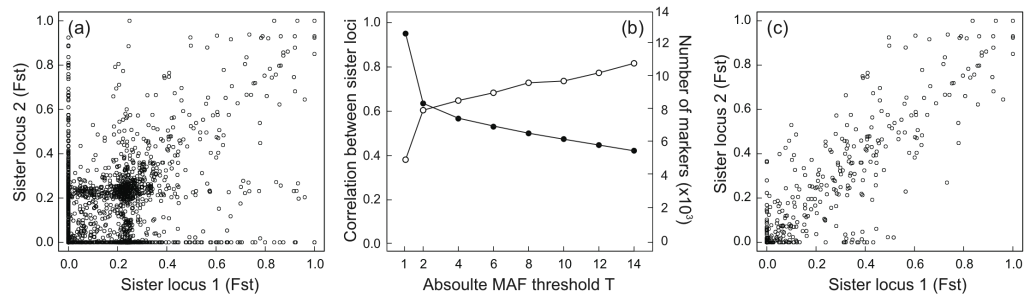


Fig. A1. Accepting SNPs without restriction on the minor allele frequency leads to a high proportion of uninformative polymorphisms in the data set, and hence to a poor correlation in F_{st} estimated by SNPs located on sister RAD loci (a). By applying increasingly restrictive minor allele frequency filters, this correlation increases rapidly (b, open dots, left axis; note that here an absolute threshold is used, i.e., independent from sample size underlying a SNP) while the number of available markers declines (b, filled dots, right axis). Imposing a minor allele frequency threshold of 0.25, the strategy chosen for our study, produces a high correlation in F_{st} between sister loci (c), indicating reasonably reliable F_{st} estimation by the SNPs.

References

Nei M, Tajima F (1981) DNA polymorphism detectable by restriction endonucleases. *Genetics* **97**, 145-163.

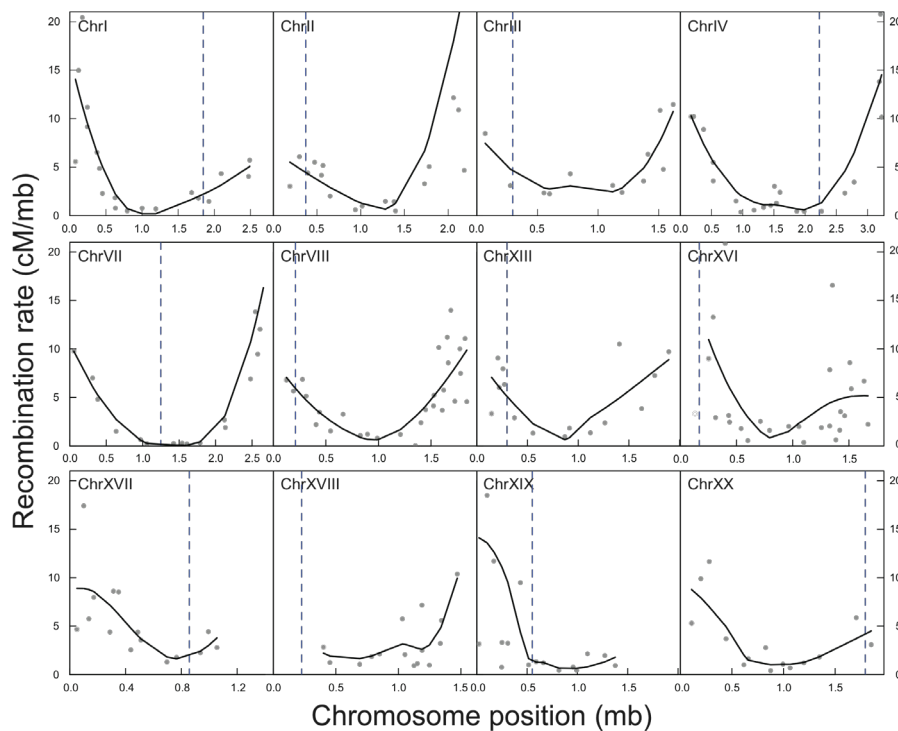
Supporting Appendix S3 to Roesti et al. 2012

Summary statistics describing the sequence data used for the study of genomic divergence between lake and stream stickleback. Values in parentheses are standard deviation, minimum, and maximum.

- Number of 76-cycle Illumina lanes: 12
- Average number of short reads per lane: 33,238,333 (3,311,888; 23,121,522; 40,385,124)
- Total number of individuals: 216
- Average number of filtered Illumina reads per individual: 1,497,549 (735,613; 30,288; 3,182,934)
- Percentage of Illumina reads successfully aligned to the stickleback genome: 71.6 (2.3; 65.4; 76.8)
- Average number of RAD loci per individual producing a consensus genotype (haploid or diploid): 38,175 (6,953; 6,546; 45,676)
- Average read coverage per consensus genotype and individual: 26.7 (12.1; 1.9; 54.7)
- Number of RAD loci with adequate genotype representation (≥ 27 consensus haplotypes) per lake-stream system. Given are the total number of loci, and the number and proportion of loci containing informative polymorphisms that were considered as markers for analysis. The latter includes only SNPs satisfying a minor variant threshold of 0.25 across the lake and stream population combined.
 - Boot: total 22,167; polymorphic 4,127 (18.6%)
 - Joe's: total 36,188; polymorphic 5,427 (15.0%)
 - Misty: total 38,062; polymorphic 8,284 (21.8%)
 - Robert's: total 36,093; polymorphic 8,417 (23.3%)

Supporting Appendix S4 to Roesti et al. 2012

Recombination rate along the chromosomes of threespine stickleback. Gray dots represent recombination rates calculated for intervals between pairs of adjacent markers underlying the linkage maps in Albert et al. (2008) and Greenwood et al. (2011). Black lines are LOESS-fits (degree: 2, bandwidth: 0.85) to these data. Dashed blue vertical lines indicate the location of the centromere, as given in Urton et al. (2011). The abscissa always spans the entire sequence of a chromosome, but in some cases data are missing for one end of the chromosome (e.g. chromosome 17). Recombination rates for nine of the 21 total chromosomes are not presented because of insufficient data (<14 data points) and/or incorrect reference chromosome assembly (or collinearity problems) suggested by directional disagreement between genetic and physical distance over large segments (chrV, IX, XII, XIV). Directional disagreement on a small scale was resolved by haphazardly eliminating closely contiguous markers. Five total outlier data points on the chromosomes 2, 7, 11, and 19 are not plotted.



References

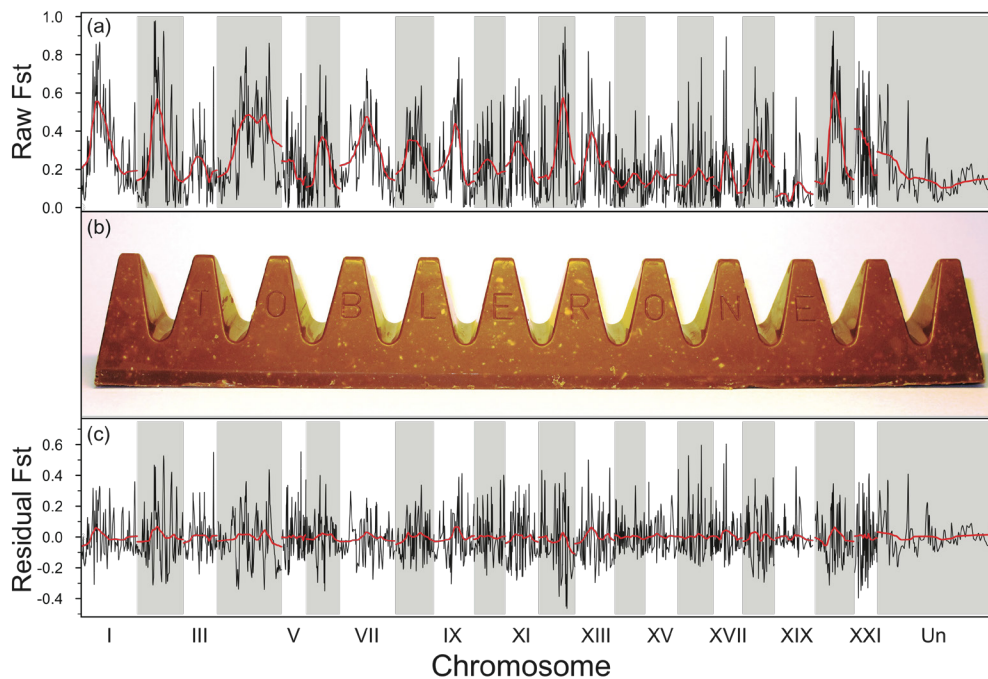
- Albert AYK, Sawaya S, Vines TH, *et al.* (2008) The genetics of adaptive shape shift in stickleback: pleiotropy and effect size. *Evolution* **62**, 76-85.
- Greenwood AK, Jones FC, Chan YF, *et al.* (2011) The genetic basis of divergent pigment patterns in juvenile threespine sticklebacks. *Heredity* **107**, 155-166.
- Urton JR, McCann SM, Peichel CL (2011) Karyotype differentiation between two stickleback species (Gasterosteidae). *Cytogenet. Genome Res.* **135**, 150-159.

Supporting Appendix S5 to Roesti et al. 2012

Statistical significance thresholds for outlier regions in sliding window analyses of lake-stream stickleback (and an allopatric lake-lake stickleback comparison) were determined through a resampling approach taking into account the correlation in divergence among SNPs due to physical linkage. Specifically, residual divergence (i.e., raw F_{st} corrected for chromosome center-biased divergence) was subjected to sliding window analysis by locally-weighted scatterplot smoothing (LOESS, moving average with bandwidth of 0.03), with physical chromosome position as predictor. The divergence predicted for every SNP in this real data set ('observed predicted divergence') was saved. Next, residual divergence was permuted 9999 times at random over the SNPs' genomic position, leading to random data sets. These new data sets were again subjected to LOESS to obtain the 'random predicted divergence' for each SNP. In each iteration, we recorded whether the random predicted divergence was greater than the observed predicted divergence. The counts across all iterations were then used to derive for each locus the probability (P-value) of observing by chance a LOESS-predicted divergence as extreme as the observed predicted divergence. Finally, we LOESS-fitted (bandwidth 0.03) the obtained P-values of the SNPs by using their observed magnitude of divergence as predictor, and screened this relationship for the magnitude of divergence required to be significant at the $P < 0.01$ level.

Supporting Appendix S6 to Roesti et al. 2012

Genome-wide divergence between lake and stream stickleback in the Boot system (a). The lines indicate the divergence predicted for each marker on all chromosomes (including the artificial 'ChrUn') by coarse (red; bandwidth = 0.3) and fine (black; bandwidth = 0.03) LOESS applied to raw F_{st} values (as in Fig. 3, top). Both line fits highlight the general tendency toward greater divergence in the chromosome centers than in their peripheries. We refer to this pattern as chromosome center-biased divergence (CCBD), emphasizing that the 'Toblerone of speciation' metaphor might be equally appropriate (b). Because CCBD is a large-scale phenomenon arising from the *combined* influence of multiple QTLs, screening F_{st} data for localized outlier signatures associated with specific QTLs benefits from a correction for CCBD (c; for details see Fig. 3).



Chapter 2

Uninformative polymorphisms bias genome scans
for signatures of selection

METHODOLOGY ARTICLE

Open Access

Uninformative polymorphisms bias genome scans for signatures of selection

Marius Roesti, Walter Salzburger and Daniel Berner*

Abstract

Background: With the establishment of high-throughput sequencing technologies and new methods for rapid and extensive single nucleotide (SNP) discovery, marker-based genome scans in search of signatures of divergent selection between populations occupying ecologically distinct environments are becoming increasingly popular.

Methods and Results: On the basis of genome-wide SNP marker data generated by RAD sequencing of lake and stream stickleback populations, we show that the outcome of such studies can be systematically biased if markers with a low minor allele frequency are included in the analysis. The reason is that these 'uninformative' polymorphisms lack the adequate potential to capture signatures of drift and hitchhiking, the focal processes in ecological genome scans. Bias associated with uninformative polymorphisms is not eliminated by just avoiding technical artifacts in the data (PCR and sequencing errors), as a high proportion of SNPs with a low minor allele frequency is a general biological feature of natural populations.

Conclusions: We suggest that uninformative markers should be excluded from genome scans based on empirical criteria derived from careful inspection of the data, and that these criteria should be reported explicitly. Together, this should increase the quality and comparability of genome scans, and hence promote our understanding of the processes driving genomic differentiation.

Keywords: Allele frequency distribution, F_{ST} , *Gasterosteus aculeatus*, Genetic marker, Hitchhiking, Population differentiation, Singleton

Background

A major challenge in evolutionary biology is to understand how natural selection acts on molecular genetic variation [1-4]. One approach to studying the consequences of selection at the genomic level is the application of genome scans that screen a collection of polymorphic genetic marker loci for their extent of differentiation between multiple (typically two) populations occupying ecologically distinct environments. Loci or genomic regions displaying particularly high population differentiation (usually quantified by an F_{ST} estimator [5]) relative to some differentiation baseline (reflecting primarily neutral drift) are interpreted as either being directly under divergent selection, or exhibiting genetic hitchhiking along with a quantitative trait locus (QTL) under divergent selection [6-9]. Genome scans therefore have the potential to illuminate the link between

ecological selection and molecular variation, and hence to contribute to our understanding of adaptive diversification. This is particularly true if information from genome scans is integrated with complementary lines of evidence such as QTL mapping [10].

Genome scans can be performed in different ways, depending on the genomic resources available for a focal research system. On the one hand, reference-free (anonymous) scans are carried out without information on the physical genomic position of a marker locus. Here the F_{ST} value for each locus is treated as an independent data point and is evaluated against a baseline distribution derived from the entire data set *e.g.*, [11-14]. Loci exhibiting extreme F_{ST} values relative to the baseline ('outlier loci') are then interpreted as being directly or indirectly influenced by divergent selection. (Note that we here use divergent selection in a broad sense, including situations where an allele is selected in one environment but neutral in the other.) On the other hand, reference-based genome scans map loci physically to an

* Correspondence: daniel.berner@unibas.ch
Zoological Institute, University of Basel, Vesalgasse 1, Basel CH-4051, Switzerland

available genome *e.g.*, [15-18]. This offers a great advantage: loci occurring in the same genomic neighbourhood, and consequently exhibiting some physical linkage, will tend to display correlated F_{ST} values that can be integrated by taking a sliding window approach. This allows not only the identification of genomic regions displaying exceptionally high population differentiation, but also exploring the number and physical extent of such regions [3]. Moreover, depending on the marker resolution, outlier regions may be screened for candidate genes potentially targeted by divergent selection.

Inferences drawn from both reference-free and reference-based genome scans obviously depend on the availability of reliable polymorphism data. The objective of our study is to highlight a potential problem with polymorphism data sets that can introduce bias to genome scans and lead to incorrect interpretations of genomic differentiation, or the lack thereof. The problem lies in F_{ST} being sensitive not only to the extent of genetic differentiation among populations, but also to the allele frequency distribution. Specifically, very low F_{ST} values (*i.e.*, near zero, or even negative values, depending on the formula used for calculation) at a polymorphic marker locus can arise for two different reasons: first, when the locus' polymorphism involves alleles segregating at relatively even frequencies in both populations, but the frequency distribution of the alleles does not differ between the populations (upper example in Table 1). For such a locus, inferring the absence of population differentiation would generally be reasonable.

Second, a very low (or negative) F_{ST} value will also arise if the alleles at a marker locus exhibit an extremely skewed frequency distribution. That is, if a locus is nearly monomorphic in both populations but contains an alternative allele segregating at very low frequency such that this allele occurs only once or a few times in the entire data set (lower example in Table 1). Such a locus is *constrained* to display a very low F_{ST} value between the populations [11]. However, inferring the absence of population differentiation from this F_{ST} value is

Table 1 Differentiation between two populations, as quantified by Weir and Cockerham's F_{ST} estimator theta [19]

	Genotypes population A			Genotypes population B			F_{ST}
	TT	TC	CC	TT	TC	CC	
Informative polymorphism	5	10	5	5	10	5	-0.026
Uninformative polymorphism	20	0	0	19	1	0	0.000

Other F_{ST} estimators produce qualitatively similar results, given informative and uninformative single nucleotide polymorphism at a marker locus (two alleles are present, T and C).

problematic. The reason is that such rare alleles primarily represent relatively recent mutations, most of which will experience rapid stochastic loss [20]. Markers with a very low minor allele frequency therefore lack the adequate sensitivity to capture the historical signatures of drift and hitchhiking, the key processes in genome scans.

To illustrate this point, imagine that a novel QTL allele arises in the neighborhood of a nearly monomorphic marker. This QTL allele is unlikely to be linked to the rare allele at the marker. If the QTL allele is favored by selection and increases in frequency within the population where it arose, hitchhiking along with the QTL will produce only a very minor (if any) allele frequency shift at the marker locus (Figure 1A). Population differentiation at the QTL will therefore not be visible at the linked marker. A clear signature of hitchhiking, however, will be seen if the marker displays a more balanced allele frequency distribution (Figure 1C; or if the QTL allele happens to be linked to the rare marker allele, Figure 1B). A similar inconsistency in differentiation between selected QTL and associated markers with highly skewed allele frequency distribution also occurs in the situation where selection acts on standing variation (soft sweep; [21]).

Of course, in addition to the situation where a *natural* allele segregates at very low frequency within populations, a highly skewed allele frequency distribution at a locus can also arise artificially during marker data acquisition. For instance due to PCR replication or sequencing error. The locus then produces a minimal F_{ST} value although correctly no F_{ST} value would be calculated because the locus is not polymorphic. However, many strategies exist to avoid such technical errors (including achieving high sequencing coverage, or re-sequencing; see also [23] and references therein). Our paper is therefore primarily concerned with biological polymorphisms.

To summarize, there are two fundamentally different causes for minimal F_{ST} values in genome scan data sets: polymorphisms with relatively even allele frequency distribution, but without population differentiation, *versus* polymorphisms with extremely skewed allele frequency distribution unable to pick up population differentiation. Hereafter, we refer to these forms of polymorphisms as 'informative' *versus* 'uninformative'. We emphasize, however, that we restrict this crude classification to genome scans searching for signatures of selection in the form of elevated differentiation. Markers with highly skewed allele frequency distributions might well be informative in other analytical contexts, such as the estimation of mutational or demographic parameters based on allele frequency spectra [24,25].

If uninformative polymorphisms are abundant in a marker data set used for a genome scan (and they

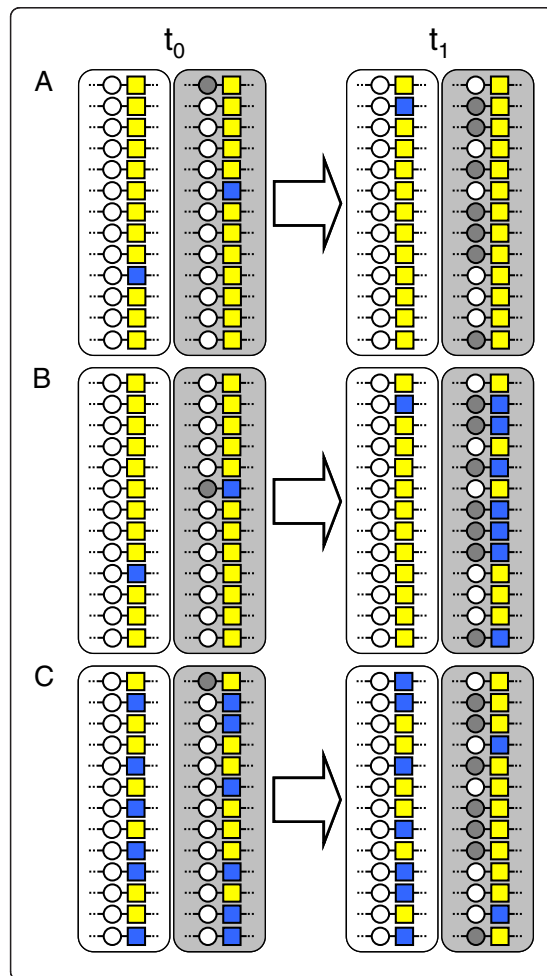


Figure 1 Informative and uninformative markers in genome scans. Two populations derived from the same ancestral population occupy ecologically distinct environments (white and gray boxes) at t_0 . Circles represent an ecologically important QTL with two alleles under divergent selection; white and gray alleles are favored in the white and gray environment. Squares represent a neutral marker with two alleles (yellow and blue). The marker is tightly physically linked to the QTL. In **A**, both initial (t_0) populations display a very low frequency for the blue marker allele. A novel adaptive QTL allele arising in the gray habitat will therefore likely be associated with the frequent yellow marker allele. When sampling the populations at t_1 , after a period of selection that has increased the frequency of the gray QTL allele in the gray environment, no signature of selection is visible at the marker locus because hitchhiking along with the favored QTL allele has not materially changed the allele frequency distribution at the marker (F_{ST} [22] approximates zero at both t_0 and t_1). In **B**, the initial conditions (t_0) are as in **A**, except that the novel adaptive QTL allele happens to be linked to the rare blue marker allele. At t_1 , selection at the QTL will be visible at the marker ($F_{ST} = 0.22$) because the blue allele has hitchhiked to high frequency. In **C**, the initial (t_0) allele frequency distribution at the marker is relatively even in both populations ($F_{ST} = 0$). At t_1 , the marker exhibits a clear signature of selection ($F_{ST} = 0.13$) because the yellow allele has increased in frequency by hitchhiking. In both **B**) and **C**) but not in **A**), we would consider the marker locus informative at t_1 based on its minor allele frequency across both samples, and consider the marker for a genome scan for the signature of selection (see text).

generally will, see below), we can predict a number of undesirable consequences: in both reference-free and reference-based approaches, the estimated overall baseline differentiation, which is considered to reflect the effect of drift, will be biased downward. As a consequence of this bias in the baseline, the number of loci considered outliers driven by divergent selection in reference-free genome scans may be inflated. By contrast, in sliding window type of scans, the magnitude of among-population differentiation in genomic regions influenced by selection will be weakened, or in the worst case erased. Both effects can lead to incorrect conclusions about the genomic consequences of divergent selection. We emphasize that these problems will arise irrespective of the specific estimator used to quantify population differentiation, or the method chosen for outlier detection. That is, uninformative marker loci will also influence sophisticated

methods that estimate F_{ST} for a locus by taking into account genome-wide differentiation and locus-specific sample size [14], or approaches based on P-values from locus-specific significance tests (e.g., [16]).

It would thus seem straightforward to eliminate uninformative marker loci from polymorphism data sets prior to performing a genome scan, as reflected in Beaumont and Nichols' [11] recommendation to preferably use loci with high heterozygosity for such analyses. However, a screen of 24 recent genome scan papers based on single nucleotide polymorphisms (SNPs), including most such studies currently available, suggests that the above issue is not generally recognized. (Note that our paper focuses on SNPs because this marker type is becoming standard in population genomics; but the conclusions hold for any type of marker.) Only three studies report marker filtering according to some minor allele frequency threshold ([18,26,27]; the latter study excluded singleton loci only, i.e., markers with the minor allele occurring only a single time). It is therefore possible that patterns reported and conclusions drawn in many genome scan studies are unreliable to some extent. Given that genome scans are becoming increasingly easy to perform owing to the advent of high-throughput sequencing technology [28], new techniques for extensive SNP discovery (in particular restriction site associated DNA (RAD) sequencing [29]), and automated data analysis pipelines, the problem of bias arising from

uninformative marker loci deserves wide recognition. A first goal of our study is therefore to use extensive SNP data from lake and stream population of stickleback fish to demonstrate that uninformative marker loci indeed have the potential to bias results from genome scans. Our second goal is to show that such bias can be avoided through careful inspection of the data set and subsequent exclusion of uninformative marker loci based on empirical criteria.

Methods

Our study uses SNP data from threespine stickleback (*Gasterosteus aculeatus*) populations occurring in lake and stream habitats within two independently colonized drainages. The first is the Lake Constance drainage in Switzerland (the 'COW' lake-stream population pair from [30]), hereafter called the 'Constance system'. The divergence between the lake and stream population in this system appears to be recent (a few hundred years). The second is the Boot Lake drainage on Vancouver Island, Canada (the Boot sites 'L' and 'S2' in [31]), hereafter called the 'Boot system'. Lake-stream divergence in this system is more ancient (thousands of years). Lake and stream stickleback are known to experience divergent selection [31,32], and the specific population pairs were chosen because they differ in the magnitude of habitat-related phenotypic and neutral genetic (microsatellite) divergence (stronger divergence in the older Boot system than in the younger Constance system). For further details on the locations and populations see [30,31]. All samples were taken with permission from the British Columbia Ministry of Environment (permit number NA06-20791), and the fisheries authority of the canton Thurgau.

For SNP detection, we Illumina-sequenced RAD [29] derived from 27 stickleback specimens from each of the four sites (*i.e.*, one lake and one stream site in two drainages; total $N = 108$). Library preparation essentially followed the method described in [17]. In short, DNA was digested by using the *SbfI* restriction enzyme and barcode-ligated for each individual separately. Amplified barcoded DNA was then single-end sequenced on an Illumina genome analyzer Ix with 76 cycles in libraries of 18 pooled individuals each. The Illumina short reads (sequenced RAD sites; deposited at the NCBI Short Read Archive, accession number SRP007695) were parsed by individual barcode, and for each individual separately aligned to the stickleback genome (Ensembl database version 63.1, assembly Broad S1) using Novoaalign v2.07.06 (<http://novocraft.com>). Alignment to a unique genome position was enforced, effectively eliminating sequences derived from repeated elements. The average sequence coverage per individual and RAD site was 27 and 31 for the lake and stream sample in the

Constance system, and 30 and 11 for the Boot system. Alignments were converted to BAM format using Samtools v0.1.11 [33]. For each individual and RAD site, we then determined the consensus diploid genotype if ten or more replicate reads were available, or a haploid consensus genotype if replication was below ten. This threshold was chosen because for polymorphic nucleotide positions, we identified heterozygote diploids based on a binomial test with insufficient power at low replication. This test involved calculating the binomial likelihood of the observed frequency distribution of the SNP alleles under the null hypothesis of heterozygosity (*i.e.*, assuming a probability of 0.5 for both alleles). Positions were considered heterozygous if the likelihood was greater than 0.01. Consensus genotyping was quality-aware in that bases with a greater than 0.01 calling error probability were excluded from the binomial test.

To find SNP markers and calculate genome-wide lake-stream population differentiation within each of the two systems, we pooled individual consensus genotypes from the lake and stream sample for each RAD site. If at least 27 genotypes were available from *each* of the two habitats, we proceeded with F_{ST} calculation. In other words, a RAD site was considered only if each individual contributed at least one haploid consensus genotype on average to the site's genotype pool. For F_{ST} calculation, the genotype pool for each RAD site was screened base by base for polymorphisms. If a variable position occurred, we calculated F_{ST} based on haplotype diversity (equation 7 in [22]). For RAD sites exhibiting multiple SNPs, we retained only the highest F_{ST} value observed across all variable base positions. (Using the average F_{ST} value across all positions, or selecting a single SNP at random, produced very similar results supporting identical conclusions.) Negative F_{ST} values were rounded to zero, as commonly done.

The above F_{ST} calculation considered *any* type of SNPs. To explore the effect of informative *versus* uninformative markers, we repeated the above F_{ST} calculation protocol by imposing the restriction that the minor (less frequent) allele had to occur at least n times in the lake-stream genotype pool, where n spanned the range from two to ten in increments of one. (The above default F_{ST} calculation represents the case with $n = 1$.) For each calculation series, we then computed the number of resulting SNPs, and the mean F_{ST} value across all SNPs. We also visualized genomic differentiation by a sliding window approach using local polynomial fitting (LOESS) implemented in R (R Development Core Team [34]; 2nd order polynomial with band width of 0.4; using simpler polynomials and different band widths did not alter our conclusions). All post-sequencing analysis except for alignment and file conversion was coded in the R

language, making use of the Bioconductor packages ShortRead [35], Biostrings, and Rsamtools.

Results

In both the Constance and Boot stickleback population pair, raising the threshold for the minimal required count of the minor SNP allele (n) had a dramatic influence on the number of polymorphic marker loci available for F_{ST} calculation. Most strikingly, the number of SNPs dropped by 46.5% (from 19,729 to 10,546) and 34% (from 16,729 to 7,546) in the Constance and Boot system when singleton loci were excluded by setting n to two (Figure 2A). Increasing n from two to ten, however, had a relatively minor effect on the number of polymorphic loci. Our stickleback data sets thus exhibit a very high proportion of singleton loci, as generally found in empirical studies (e.g., [36-39]). The genomic location of these singleton loci did not show any systematic association with chromosome position (details not).

Including these uninformative marker loci in the genome scan led to the consequences predicted above. First, baseline differentiation was substantially lower than the differentiation obtained when setting n to two or greater (Figure 2B). For instance, genome-wide F_{ST} increased by 17% and 20% in the Constance and Boot system when raising n from one to two. In absolute terms, this shift was more dramatic in the Boot system displaying the higher overall differentiation between the populations. Second, F_{ST} profiles obtained from sliding window

analyses including all markers ($n = 1$) were strikingly flatter than those from analyses excluding uninformative polymorphisms. These two consequences are visualized for a segment of chromosome seven (Figure 3), which is representative of what we found throughout the genome. For that specific genomic region, analyses with and without uninformative marker loci might lead to qualitatively different conclusions about the magnitude and physical extent of population differentiation. For example, in the Constance system, a large segment ranging approximately from 12–14 mb displays elevated differentiation, as revealed when using informative markers only. This differentiation is certainly substantial, given the low baseline differentiation in that young system (Figure 2B), and might indicate ongoing divergent selection in that genomic region. Nevertheless, elevated differentiation within that region would probably not be recognized when tolerating uninformative markers in the sliding window analysis.

Note that in Figure 3, we define informative marker loci as those with the minor allele occurring at least four times ($n = 4$), resulting in an average inter-locus distance of 53 kb and 63 kb for the Constance and Boot system. This minor allele threshold eliminated bias associated with uninformative marker loci relatively effectively; choosing higher thresholds had a relatively minor influence on the sliding window profiles.

Discussion

Our empirical analysis demonstrates that abundant uninformative polymorphisms in a genome scan data set can bias the estimated baseline differentiation, and hence affect conclusions about the genomic signatures of selection.

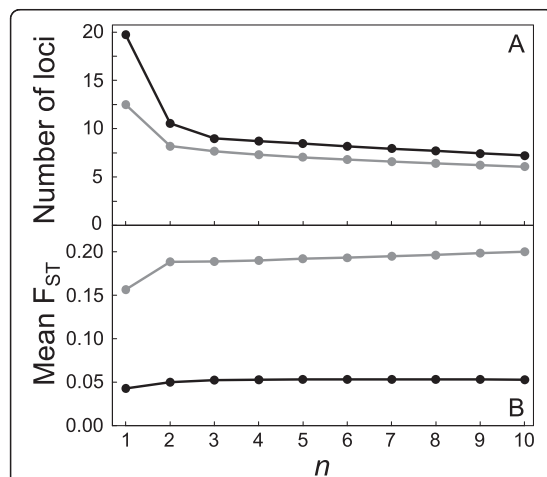


Figure 2 The number of polymorphic loci ($\times 10^3$) (A), and mean F_{ST} across all loci (B), for different minor allele count thresholds (n) in the Constance (black) and Boot (gray) lake-stream stickleback system. This threshold specifies the minimum number of times the minor SNP allele at a locus had to occur in the pooled lake and stream sample for a polymorphic locus to remain in the data set.

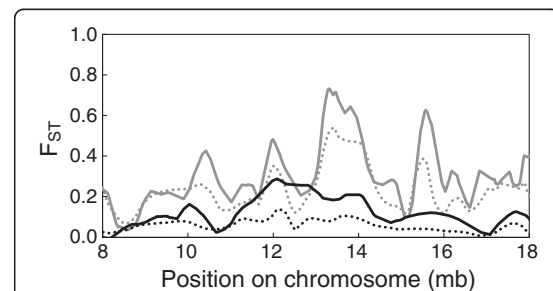


Figure 3 Differentiation along a segment of chromosome seven between the lake and stream stickleback population from the Constance (black) and Boot (gray) system. Sliding window analysis was performed by local polynomial fitting of F_{ST} values for data sets with the allele frequency threshold n set to one (all SNPs in the data sets considered; dotted lines), and n set to four (at least four copies of the minor allele required across the pooled lake and stream sample in each system; solid lines). Note the relatively flat differentiation profiles with $n = 1$.

In our stickleback data set, uninformative polymorphisms (essentially in the form of singleton loci) were very abundant. Illumina sequencing type one errors (*i.e.*, wrong base calls despite high indicated base call quality) in RAD sequences poorly replicated at the individual level might contribute to this pattern [23,39]. To examine this possibility, we inspected 50 randomly chosen SNPs exhibiting zero F_{ST} from the full data set accepting any type of polymorphisms (*i.e.*, minor allele count threshold $n=1$) for each lake-stream system. As expected, a high proportion of these markers were singleton loci (Constance: 28 [56%]; Boot: 35 [70%]). For the Boot system with lower replication per locus in the stream sample (see above), 15 of the 35 singleton loci represented unreplicated RAD sequences. For these loci, the minor allele is likely a sequencing error.

However, all but one of the singleton alleles in the Constance system represented consensus genotypes integrating multiple (2–26, mean: 9.1) replicate RAD sequences. Hence, the bulk of the uninformative marker loci in our data clearly *cannot* be attributed to sequencing error, because the probability of multiple identical errors at a specific nucleotide position at a given RAD site is practically zero. The abundance of rare SNP alleles therefore represents a real biological feature of the studied stickleback populations (acknowledging a small potential contribution from PCR artefacts). This is not unexpected: theory consistently predicts a skew toward polymorphisms with low minor allele frequency, and hence a high proportion of singleton polymorphisms, under a broad range of demographic and selective conditions [24,36,40–44]. Bias associated with uninformative polymorphisms is therefore of general importance to genome scan studies, and not specific to our empirical system. Our analysis also raises a caveat regarding marker densities; the effective number of markers providing relevant information in genome scans might often be dramatically lower than the number reported.

In the present study, excluding singleton polymorphisms had the greatest influence on the results. Reliable quantification of differentiation patterns, however, might require substantially more stringent minor allele frequency thresholds. (Note that such marker filtering also effectively eliminates any sequencing and PCR error from the data.) Bradbury *et al.* [27], for instance, excluded SNPs exhibiting an overall minor allele frequency of 0.25 or less, and a similar threshold was adopted in a recent lake-stream stickleback study carried out in our lab [45]. To obtain a guideline for marker filtering, the latter RAD-based study evaluated the strength of the correlation in F_{ST} values between 'sister' RAD sites (*i.e.*, DNA sequences flanking the *same* restriction site in the genome) in relation to increasingly

stringent minor allele frequency thresholds (see Appendix S2 in the Supporting information to [45]). The rationale was that if an F_{ST} value provided by a given marker reliably quantifies the consequences of drift and selection in a genomic region, then another extremely tightly linked marker should yield a similar F_{ST} value. This approach, however, requires tightly physically linked markers data and substantial population differentiation (otherwise the correlation in F_{ST} between linked will remain poor even with stringent marker filtering).

Conclusions

Given the rapidly increasing feasibility and popularity of genome scans for signatures of selection, researchers should be aware that uninformative polymorphisms need to be excluded from data sets. This is not achieved by just avoiding technical errors, as a high prevalence of nearly monomorphic loci is a general biological feature of samples from natural populations. We suggest that a reasonable strategy to define and eliminate uninformative polymorphisms should be chosen by inspecting the allele frequency distribution of the polymorphisms, and by assessing the influence of different marker filtering thresholds on the genomic patterns of interest, or appropriate statistics (such as the correlation of F_{ST} between sister RAD sites). Also, the approach taken to eliminate uninformative polymorphisms should be reported explicitly. Together, this should increase the quality and comparability of genome scans, and hence promote our understanding of the processes shaping genomic differentiation.

Competing interest

The authors declare no competing interest.

Acknowledgment

Field work was aided by Anne-Catherine Grandchamp and, for the Boot system, supported financially by Andrew Hendry. Roman Kistler (fisheries authority of the canton Thurgau) permitted sampling of the Constance specimens. Paul Etter and Bill Cresko kindly shared their experience and protocol for RAD library preparation. Brigitte Aeschbach and Nicolas Boileau facilitated wet lab work, Ina Nissen performed Illumina sequencing at the Quantitative Genomics Facility, D-BSSE, ETH Zürich, and Lukas Zimmermann provided IT support. The bioinformatics pipeline benefited from modifications to Novoalign by Colin Hercus, and from coding suggestions by Martin Morgan and Hervé Pagès. Matthieu Foll, Markus Pfenninger, and three anonymous reviewers provided valuable suggestions that improved the paper. WS was supported financially by the European Research Council (Starting Grant 'INTERGENADAPT'), the Swiss National Science Foundation, and the University of Basel. DB was supported by the Swiss National Science Foundation (Ambizione grant PZ00P3_126391 / 1) and the Research Foundation of the University of Basel. We kindly thank all these people and institutions.

Authors' contributions

DB and MR conceived the study; WS provided materials and infrastructure; MR and DB generated the sequence data; DB and MR analyzed the data; DB wrote the paper, with input from MR and WS. All authors read and approved the final manuscript.

Received: 24 January 2012 Accepted: 22 June 2012

Published: 22 June 2012

References

1. Wu CI: The genic view of the process of speciation. *J Evol Biol* 2001, **14**:851–865.
2. Mitchell-Olds T, Willis JH, Goldstein DB: Which evolutionary processes influence natural genetic variation for phenotypic traits? *Nat Rev Genet* 2007, **8**:845–856.
3. Nosil P, Funk DJ, Ortiz-Barrientos D: Divergent selection and heterogeneous genomic divergence. *Mol Ecol* 2009, **18**:375–402.
4. Schluter D: Evidence for ecological speciation and its alternative. *Science* 2009, **323**:737–741.
5. Holsinger KE, Weir BS: Genetics in geographically structured populations: defining, estimating and interpreting F_{ST} . *Nat Rev Genet* 2009, **10**:639–650.
6. Lewontin RC, Krakauer J: Distribution of gene frequency as a test of theory of selective neutrality of polymorphisms. *Genetics* 1973, **74**:175–195.
7. Maynard Smith J, Haigh J: Hitch-hiking effect of a favorable gene. *Genet Res* 1974, **23**:23–35.
8. Beaumont MA: Adaptation and speciation: what can F_{ST} tell us? *Trends Ecol Evol* 2005, **20**:435–440.
9. Storz JF: Using genome scans of DNA polymorphism to infer adaptive population divergence. *Mol Ecol* 2005, **14**:671–688.
10. Stinchcombe JR, Hoekstra HE: Combining population genomics and quantitative genetics: finding the genes underlying ecologically important traits. *Heredity* 2008, **100**:158–170.
11. Beaumont MA, Nichols RA: Evaluating loci for use in the genetic analysis of population structure. *Proc R Soc Lond B* 1996, **263**:1619–1626.
12. Beaumont MA, Balding DJ: Identifying adaptive genetic divergence among populations from genome scans. *Mol Ecol* 2004, **13**:969–980.
13. Foll M, Gaggiotti O: A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. *Genetics* 2008, **180**:977–993.
14. Excoffier L, Hofer T, Foll M: Detecting loci under selection in a hierarchically structured population. *Heredity* 2009, **103**:285–298.
15. Akey JM, Zhang G, Zhang K, Jin L, Shriver MD: Interrogating a high-density SNP map for signatures of natural selection. *Genome Res* 2002, **12**:1805–1814.
16. Turner TL, Hahn MW, Nuzhdin SV: Genomic islands of speciation in *Anopheles gambiae*. *PLoS Biol* 2005, **3**:e285.
17. Hohenlohe PA, Bassham S, Etter PD, Stiffler N, Johnson EA, Cresko WA: Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLoS Genet* 2010, **6**:e1000862.
18. Lawnczak MKN, Emrich SJ, Holloway AK, Regier AP, Olson M, White B, Redmond S, Fulton L, Appelbaum E, Godfrey J, et al: Widespread divergence between incipient *Anopheles gambiae* species revealed by whole genome sequences. *Science* 2010, **330**:512–514.
19. Weir BS, Cockerham CC: Estimating F-statistics for the analysis of population-structure. *Evolution* 1984, **38**:1358–1370.
20. Ewens W: *Mathematical population genetics*. New York: Springer; 1979.
21. Hermisson J, Pennings PS: Soft sweeps: molecular population genetics of adaptation from standing genetic variation. *Genetics* 2005, **169**:2335–2352.
22. Nei M, Tajima F: DNA polymorphism detectable by restriction endonucleases. *Genetics* 1981, **97**:145–163.
23. Nielsen R, Paul JS, Albrechtsen A, Song YS: Genotype and SNP calling from next-generation sequencing data. *Nat Rev Genet* 2011, **12**:443–451.
24. Marth GT, Czabarka E, Murvai J, Sherry ST: The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. *Genetics* 2004, **166**:351–372.
25. Keightley PD, Eyre-Walker A: Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies. *Genetics* 2007, **177**:2251–2261.
26. Oleksyk TK, Zhao K, De La Vega FM, Gilbert DA, O'Brien SJ, Smith MW: Identifying selected regions from heterozygosity and divergence using a light-coverage genomic dataset from two human populations. *PLoS One* 2008, **3**:e1712.
27. Bradbury IR, Hubert S, Higgins B, Borza T, Bowman S, Paterson IG, Snelgrove PVR, Morris CJ, Gregory RS, Hardie DC, et al: Parallel adaptive evolution of Atlantic cod on both sides of the Atlantic Ocean in response to temperature. *Proc R Soc B* 2010, **277**:3725–3734.
28. Mardis ER: The impact of next-generation sequencing technology on genetics. *Trends Genet* 2008, **24**:133–141.
29. Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, Selker EU, Cresko WA, Johnson EA: Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS One* 2008, **3**:e3376.
30. Berner D, Roesti M, Hendry AP, Salzburger W: Constraints on speciation suggested by comparing lake-stream stickleback divergence across two continents. *Mol Ecol* 2010, **19**:4963–4978.
31. Berner D, Grandchamp A-C, Hendry AP: Variable progress toward ecological speciation in parapatry: stickleback across eight lake-stream transitions. *Evolution* 2009, **63**:1740–1753.
32. Berner D, Adams DC, Grandchamp A-C, Hendry AP: Natural selection drives patterns of lake-stream divergence in stickleback foraging morphology. *J Evol Biol* 2008, **21**:1653–1665.
33. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R: Genome Project Data P: The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009, **25**:2078–2079.
34. R Development Team: *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing; 2010.
35. Morgan M, Anders S, Lawrence M, Aboyoun P, Pagès H, Gentleman R: ShortRead: a bioconductor package for input, quality assessment and exploration of high-throughput sequence data. *Bioinformatics* 2009, **25**:2607–2608.
36. Harpending HC, Batzer MA, Gurven M, Jorde LB, Rogers AR, Sherry ST: Genetic traces of ancient demography. *Proc Natl Acad Sci USA* 1998, **95**:1961–1967.
37. Barreiro LB, Laval G, Quach H, Patin E, Quintana-Murci L: Natural selection has driven population differentiation in modern humans. *Nat Genet* 2008, **40**:340–345.
38. Geiler KA, Harrison RG: A $\Delta 11$ desaturase gene genealogy reveals two divergent allelic classes within the European corn borer (*Ostrinia nubilalis*). *BMC Evol Biol* 2010, **10**:112.
39. Keightley PD, Halligan DL: Inference of site frequency spectra from high-throughput sequence data: quantification of selection on nonsynonymous and synonymous sites in humans. *Genetics* 2011, **188**:931–940.
40. Nei M, Li W-H: The transient distribution of allele frequencies under mutation pressure. *Genet Res* 1976, **28**:205–214.
41. Li WH: Maintenance of genetic variability under joint effect of mutation, selection and random drift. *Genetics* 1978, **90**:349–382.
42. Fu YX: Statistical properties of segregating sites. *Theor Pop Biol* 1995, **48**:172–197.
43. Eberle MA, Kruglyak L: An analysis of strategies for discovery of single-nucleotide polymorphisms. *Genet Epidemiol* 2000, **19**:29–35.
44. Evans SN, Shvets Y, Slatkin M: Non-equilibrium theory of the allele frequency spectrum. *Theor Pop Biol* 2007, **71**:109–119.
45. Roesti M, Hendry AP, Salzburger W, Berner D: Genome divergence during evolutionary diversification as revealed in replicate lake-stream stickleback population pairs. *Mol Ecol* 2012, **21**:2852–2862.

doi:10.1186/1471-2148-12-94

Cite this article as: Roesti et al.: Uninformative polymorphisms bias genome scans for signatures of selection. *BMC Evolutionary Biology* 2012 **12**:94.

Submit your next manuscript to BioMed Central and take full advantage of:

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at
www.biomedcentral.com/submit



Chapter 3

Recombination in the threespine stickleback genome—patterns and consequences

Recombination in the threespine stickleback genome—patterns and consequences

MARIUS ROESTI, DARIO MOSER and DANIEL BERNER

Zoological Institute, University of Basel, Vesalgasse 1, CH-4051 Basel, Switzerland

Abstract

Heterogeneity in recombination rate may strongly influence genome evolution and entail methodological challenges to genomic investigation. Nevertheless, a solid understanding of these issues awaits detailed information across a broad range of taxa. Based on 282 F_2 individuals and 1872 single nucleotide polymorphisms, we characterize recombination in the threespine stickleback fish genome. We find an average genome-wide recombination rate of 3.11 cM/Mb. Crossover frequencies are dramatically elevated in the chromosome peripheries as compared to the centres, and are consistent with one obligate crossover per chromosome (but not chromosome arm). Along the sex chromosome, we show that recombination is restricted to a small pseudoautosomal domain of *c.* 2 Mb, spanning only 10% of that chromosome. Comparing female to male RAD sequence coverage allows us to identify two discrete levels of degeneration on the Y chromosome, one of these 'evolutionary strata' coinciding with a previously inferred inverted region. Using polymorphism data from two young (<10 000 years old) ecologically diverged lake-stream population pairs, we demonstrate that recombination rate correlates with both the magnitude of allele frequency shifts between populations and levels of genetic diversity within populations. These associations reflect genome-wide heterogeneity in the influence of selection on linked sites. We further find a strong relationship between recombination rate and GC content, possibly driven by GC-biased gene conversion. Overall, we highlight that heterogeneity in recombination rate has profound consequences on genome evolution and deserves wider recognition in marker-based genomic analyses.

Keywords: evolutionary strata, *Gasterosteus aculeatus*, GC content, genetic diversity, pseudoautosomal region, RAD sequencing, sex chromosome

Received 2 March 2013; accepted 17 March 2013

Introduction

Meiotic recombination is a fascinating process because of its pivotal role in multiple biological contexts. For instance, recombination is generally considered instrumental to the proper segregation of homologous chromosomes during meiosis (Mather 1938; Baker *et al.* 1976; Roeder 1997; Smith & Nicolas 1998; Hassold & Hunt 2001). At the same time, recombination breaks the linkage between DNA segments located on the same chromosome. This allows selection to operate more effectively on multiple loci, and hence promotes adaptation

(Hill & Robertson 1966; Felsenstein 1974; Otto & Barton 1997; Burt 2000; Otto & Lenormand 2002). Conversely, the suppression of recombination can initiate chromosome degeneration, a process believed to be common during sex chromosome evolution (Bull 1983; Charlesworth & Charlesworth 2000; Charlesworth *et al.* 2005; Wilson & Makova 2009).

Variation in recombination rate may also explain genome-wide heterogeneity in the magnitude of genetic divergence between populations, and genetic diversity within populations. The reason is that linkage between selected loci and their physical neighbourhood is tighter in regions exhibiting relatively low recombination rate. Selectively neutral polymorphisms will therefore be affected by selection more often and more strongly

Correspondence: Daniel Berner, Fax: +41 (0) 61 267 0301; E-mail: daniel.berner@unibas.ch

when located in low-recombination regions. As a consequence, hitchhiking under positive and background selection is predicted to increase allele frequency shifts between populations, and to reduce genetic diversity within populations, in low-recombination to high-recombination genomic regions (Maynard Smith & Haigh 1974; Kaplan *et al.* 1989; Begun & Aquadro 1992; Nordborg *et al.* 1996; Charlesworth *et al.* 1997; Charlesworth 1998; Nachman 2002). Similarly, loci under divergent selection between ecologically distinct habitats should impede neutral gene flow more extensively in low-recombination regions (Barton & Bengtsson 1986; Feder & Nosil 2010). Finally, recombination may have direct effects on the constitution of chromosomes, for instance through biased gene conversion or mutagenesis (Galtier *et al.* 2001; Duret & Galtier 2009; Webster & Hurst 2012).

Despite the recognition of recombination as a major evolutionary factor, our understanding of both the mechanisms governing the process, and its consequences on genome evolution, remains highly incomplete (Nachman 2002; Smukowski & Noor 2011; Webster & Hurst 2012). Moreover, detailed investigations of recombination outside genetic model organisms are needed for the discovery of general patterns. The goal of our study is to provide the first comprehensive analysis of meiotic recombination in threespine stickleback fish (*Gasterosteus aculeatus*).

A thorough understanding of recombination in this powerful model organism for ecological genetics is particularly valuable for two reasons. First, the species has been shown to display a relatively young (<10 Myr old) XY (male-heterogametic) sex determination system (Peichel *et al.* 2004). Information on the extent of XY recombination and associated patterns of Y degeneration, however, remains highly incomplete, but promises exciting insights into sex chromosome evolution (Peichel *et al.* 2004; Ross & Peichel 2008; Shikano *et al.* 2011). Second, performing genome scans in stickleback populations residing contiguously in selectively distinct lake and stream habitats, we have shown recently that population divergence (F_{ST}) is elevated in chromosome centres and argued that this effect is caused by a lower recombination rate within these regions (Roesti *et al.* 2012a). Because robust information on recombination was lacking, however, this hypothesis could not be evaluated definitively. Our study therefore combines single nucleotide polymorphism (SNP) data from a laboratory F_2 cross and natural populations to characterize the stickleback recombination landscape; to explore the role of recombination in sex chromosome evolution; to examine the relationship between recombination rate and the magnitude of divergence among and genetic diversity within populations; and to investigate the

association between recombination rate and nucleotide composition.

Materials and methods

Laboratory cross

We generated an F_2 population for linkage map construction by artificially crossing a male and a female from the Central European ROM and CHE populations (described in Berner *et al.* 2010; Moser *et al.* 2012; Roesti *et al.* 2012b) in the spring 2009. The resulting F_1 were raised in two 50-L tanks on a mixed *Artemia* (live, decapsulated cysts, frozen) and bloodworm diet under 'summer' laboratory conditions (18–20 °C with a 16:8-h day/night photoperiod). After a 'winter' phase (15 °C, 8:16-h photoperiod) of 3 months, summer conditions were re-established in the spring 2010 to initiate reproduction. The F_2 population was generated by performing 20 artificial F_1 full-sib crosses, each involving a unique male–female combination. After 1 year, 282 adult F_2 (140 males, 142 females) were haphazardly chosen, killed with an overdose of MS-222 and stored in absolute ethanol.

Marker generation

DNA was extracted from pectoral fin tissue on a MagNA Pure LC278 extraction robot (Roche) by using the tissue Isolation Kit II. We then prepared restriction site-associated DNA (RAD; Baird *et al.* 2008) libraries, involving *Sbf1* restriction, the fusion of 5-mer individual barcodes and pooling DNA of 62 individuals per library. The final enrichment PCR was performed in duplicate to reduce random amplification variation. Each library was single-end sequenced to 100 base reads in a separate Illumina HiSeq lane. In addition to the F_2 individuals, the two founder individuals of the cross were also sequenced, each twice in different libraries.

The Illumina sequences were sorted according to barcode and aligned to the stickleback reference genome (release Broad S1; Jones *et al.* 2012) by using Novoalign v2.07.06 (<http://novocraft.com>), accepting a total of approximately eight high-quality mismatches and/or indels along a read. Alignments were converted to BAM format using SAMtools (Li *et al.* 2009). Each replicate alignment of each grandparent was then screened independently for homozygous RAD loci. A locus qualified as homozygous if it was either invariant or if the binomial probability for the two dominant haplotypes to reflect a heterozygous locus was <0.001. We here ignored loci with <12 \times coverage (average coverage per locus varied between 31 and 47 among the grandparents

and replicate alignments). RAD loci proving homozygous in *both* replicates of a given grandparent were then screened for SNPs (here subsuming both SNPs and microindels) fixed for different alleles between the grandparents, accepting only one SNP per RAD locus. This conservative SNP detection strategy yielded a total of 2223 markers.

The F₂ population was then genotyped at each SNP detected in the grandparents. We considered a SNP homozygous when only one grandparent allele was present and occurred in at least 20 copies, or heterozygous when both alleles were present in at least 20 copies each (average sequence coverage per RAD locus was 55.6 among the F₂ individuals). Loci not satisfying these criteria received an ambiguous genotype based on the dominant allele or were treated as missing data when the total allele count was below six (<0.5% of all genotypes). Next, we eliminated 58 SNPs displaying clearly skewed allele frequencies across the F₂ and ordered the remaining 2165 markers according to their physical position in the Broad S1 stickleback reference genome.

Genome reassembly

Visual inspection of the genotypes ordered according to the reference genome indicated marker intervals exhibiting extremely high crossover frequency. Without exception, these intervals coincided perfectly with scaffold boundaries, indicating genome assembly errors. This conclusion was also supported by comparing physical and genetic map positions in a low-resolution data set extracted by Roesti *et al.* (2012a) from genetic maps available for North American stickleback (Albert *et al.* 2008; Greenwood *et al.* 2011): markers on scaffolds found to be inverted in the current study showed opposite genetic and physical map order in the latter data set as well (details not presented).

An accurate characterization of recombination thus required genome reassembly. For this, we created *de novo* linkage groups in R/qtl (Broman & Sen 2009) by including markers unanchored to any linkage group in the Broad S1 genome. We used a maximum recombination frequency of 0.3 or less and a LOD score of 8 or greater and further optimized marker order along linkage groups through permutation within a sliding window of seven markers. The resulting genetic map allowed us to invert 13 total scaffolds (size range: 0.7–17.1 Mb; 98.2 Mb in total) within the known linkage groups (hereafter ‘chromosomes’) and to incorporate 18 previously unanchored scaffolds with a total length of 20.1 Mb into the chromosomes. We ignored unanchored scaffolds smaller than 140 kb, as this was below our average marker resolution. We then recalculated the physical position for every marker. These assembly

corrections are described in Appendix S1 (Supporting information). All physical map positions in this study refer to our improved genome assembly, which is available in FASTA format on the Dryad digital repository (doi:10.5061/dryad.846nj).

For final genetic map construction, we first corrected genotyping errors and ambiguous calls by hand, making the common and well-supported assumption that the vast majority of tight double-recombinants reflect genotyping errors. We then clipped the most peripheral marker at each chromosome end because here phase shifts were most difficult to distinguish from genotyping errors. Next, we discarded all markers not assigned to linkage groups, and one (redundant) marker in cases where two SNPs formed a pair derived from sister RAD loci (i.e. loci flanking the same *Sbfl* restriction site). The final data set used for genetic mapping comprised 1872 markers (59–150 per chromosome), with an average spacing of 217 kb. The genotype data are provided as Appendix S2 (autosomes) and S3 (sex chromosome, Supporting information), and linkage map and corresponding physical map positions (the latter before and after reassembly) are listed in Appendix S4 (Supporting information). We note that this data set is expected to slightly underestimate recombination rate along chromosomes. The reason is that with our marker resolution, a few tight double-crossovers may have escaped detection altogether, and a few others may have been captured by one or two markers but taken as genotyping error and eliminated. Moreover, our markers never covered the full physical chromosome span because of the randomness of *Sbfl* restriction sites; because we ignored unanchored scaffolds mapping to one or both ends of many chromosomes when these scaffolds were small and represented by only one to three markers; and because we discarded the peripheral marker on each end of the initially generated linkage groups.

Analysis of recombination

Genetic distances along the 20 autosomes were estimated in R/qtl using the Kosambi map function (assuming crossover interference) and the full F₂ panel. For the sex chromosome (chromosome 19; Peichel *et al.* 2004), final map construction used genotype data from females only ($N = 142$). The reason is that sequence degeneration of the Y relative to the X chromosome precluded reliable genotyping in males (the reference sex chromosome sequence is the X). R/qtl was also used to count the number of crossovers for each individual and chromosome.

We visualized recombination rate along the chromosomes by plotting genetic distance (cM) against physical

distance (Mb). Moreover, we calculated the average recombination rate for every interval between adjacent markers as the ratio of genetic distance to physical distance (cM/Mb) and plotted this rate against the physical midpoint of the marker interval. We also calculated average recombination rate across each chromosome, and across each chromosome arm, using for the latter information on centromere positions extracted from Urton *et al.* (2011). Throughout this paper, effective physical chromosome (and chromosome arm) spans are defined by the position of our most peripheral markers. The only exception is Appendix S1 (Supporting information) where we show the full physical chromosome lengths.

Crossover counts were used to examine the relationship between recombination frequency and chromosome length. We here determined for each chromosome the average crossover number across the 282 F₂ (or the 142 females for chromosome 19) and calculated the correlation coefficient *r* between this variable and chromosome length. The magnitude of this test statistic was evaluated against its empirical random distribution established by permuting the crossover data 9999 times (Manly 2007; all statistical tests in this study are based on analogous permutation tests). A similar analysis was performed by using chromosome arm length, rather than total chromosome length, as a predictor of crossover number. In this latter analysis, six chromosome arms with low marker coverage were excluded, which had a trivial influence on the results. Also, these analyses were performed with and without the sex chromosome. As this did not materially influence the results, we report the former.

Individual crossover counts across all autosomes were used to test for a difference in overall recombination rate between the sexes, using as test statistic the F-ratio of a linear model with crossover count as response and sex as fixed factor. Analogous tests were also performed to explore sex differences in crossover number for each chromosome separately. Finally, individual crossover counts were used to scan the genome (including chromosome 19) for the presence of quantitative trait loci (QTL) determining recombination rate. We emphasize that our data are not ideal for this purpose; quantifying the recombination phenotype in the F₂ generation would have required crossover data from the F₃ generation or from F₂ gametes. Our scan was thus limited to detecting QTL heterozygous in one or both of the grandparents. The QTL scan was performed in R/qtl using the extended Haley–Knott method (other methods produced very similar results). Significance of LOD peaks was established based on 9999 permutations, following Broman & Sen (2009).

Recombination and divergence within the sex chromosome

Recombination between the X and Y chromosomes was studied by determining which of the 69 SNPs along chromosome 19 occurred homozygous for the grandfather allele in F₂ females. This female genotype necessarily requires XY crossover in the F₁ father.

To explore degeneration of the Y chromosome, we haphazardly selected 100 males and 100 females from the F₂ population. For each sex separately, we determined for every RAD locus along chromosome 19 the total sequence coverage across all individual alignments. For each RAD locus, we then calculated the ratio of female to male coverage. A RAD locus not or little differentiated between the gametologs would display an expected ratio of one because both the X and Y sequences would align to the X reference. At a locus substantially diverged between X and Y, the latter would no longer align to the reference, producing twice the sequence coverage in females relative to males. To reduce noise, we restricted this analysis to RAD loci displaying a minimal total sequence coverage of 4000 in each sex, yielding a total of 1556 informative loci along the X chromosome (average intermarker distance: 13 kb). This analysis of Y degeneration was additionally performed by using a natural population sample from Europe (CHE) and Canada (Boot Lake, see below). These populations are derived independently from Atlantic and Pacific ancestors. Because here sample size was much smaller (*N* = 13–14 per sex and population), we used a minimal sequence coverage threshold of 50 per sex.

Genetic divergence, genetic diversity and GC content in relation to recombination rate

We tested the prediction of a negative genome-wide correlation between recombination rate and the magnitude of allele frequency shifts by using divergence data from two independent replicate lake–stream population pairs studied in Roesti *et al.* (2012a) (the Boot and Robert's pair; see also Berner *et al.* 2008, 2009). These young (postglacial, <10 000 years old) population pairs are those among the four pairs investigated in Roesti *et al.* (2012a) displaying the strongest divergence in phenotypes and genetic markers between the selectively distinct habitats (genome-wide median *F*_{ST} is 0.15 and 0.03 for Boot and Robert's; Roesti *et al.* 2012a). Each of the four samples was represented by 27 individuals (balanced sex ratio). Polymorphism data were generated through RAD sequencing, yielding 3930 and 7992 genome-wide SNPs for the Boot and Robert's pair (details on library preparation, sequencing, genotyping,

SNP detection and access to the raw data are given in Roesti *et al.* 2012a,b). The magnitude of divergence between the lake and stream population was quantified by F_{ST} based on haplotype diversity (Nei & Tajima 1981; formula 7), tolerating only informative SNPs with a minor allele frequency of 0.25 or greater (Roesti *et al.* 2012b). F_{ST} was then averaged across the intervals defined by adjacent markers from the mapping cross, resulting in the same resolution as our recombination rate data (see Fig. 1). This allowed us to explore the genome-wide correlation between the magnitude of divergence and recombination rate, using r as statistic for significance testing.

Next, we examined the prediction of a positive correlation between recombination rate and levels of genetic diversity within each population. For this, we screened each of the four population samples separately for polymorphisms and calculated genetic diversity (haplotype diversity, Nei & Tajima 1981; singletons were omitted to exclude technical artefacts). RAD loci were allowed to contribute a single SNP only, keeping the one with the highest diversity at loci with multiple SNPs (drawing a SNP at random produced very similar results). The resulting total number of SNPs varied between 4938 and 17 649 among the populations. As a complementary analysis, we also counted the number of polymorphisms (excluding singletons) on each RAD locus, with the number of RAD loci varying between 6440 and 25 186 among the populations. We considered these data, hereafter referred to as SNP density, a valuable alternative genetic diversity metric because selection should not only skew allele frequencies in linked regions, but also reduce the density of polymorphisms in those regions. Both the genetic diversity and SNP density data were averaged to the resolution of the genetic map and tested for an association with recombination rate as described for F_{ST} .

Finally, we investigated a possible association between recombination rate and GC content in an analogous way. However, to maximize precision, we calculated the proportion of GC nucleotides for each marker interval based on the full reference genome sequence rather than our RAD sequences. Moreover, we here detected a clear nonlinear relationship and therefore used as test statistic the ratio of residual to total sum of squares of a nonparametric regression (LOESS—robust locally weighted scatterplot smoothing; Cleveland 1979; a linear fit with r as test statistic produced similar results). We note that this analysis assumes that patterns of nucleotide composition in the reference genome, which was built based on a Pacific-derived freshwater stickleback, are also representative of Atlantic-derived European populations. This assumption is justified; repeating the correlation analysis using

genome-wide GC content estimated from consensus sequences at 27 396 RAD loci derived from the cross grandmother produced similar results (details not presented).

In the above correlation analyses (F_{ST} , genetic diversity, SNP density, GC content), marker intervals with an extreme recombination rate (below 0.01 and above 40) were excluded, although analyses including *all* intervals produced very similar results. The final data sets thus comprised 1783 genome-wide marker intervals. Also, including or excluding the sex chromosome did not materially influence the analyses; we thus report the former. Apart from sequence alignment and BAM conversion, all analyses and plotting were carried out in the R language (R Development Core Team 2012), benefiting greatly from the Bioconductor packages ShortRead (Morgan *et al.* 2009), Rsamtools and Biostrings. Data smoothing was performed with R's implementation of LOESS.

Results

The 21 stickleback chromosomes accounted for a total genetic map length of 1251 cM, yielding a genome-wide average recombination rate of 3.11 cM/Mb (this number is based on the total physical genome length effectively covered by our markers: 401.8 Mb). However, recombination rate proved highly heterogeneous across the genome: crossovers occurred primarily in the chromosome peripheries, with a greatly reduced rate in the chromosome centres (Fig. 1). Except for two of the smallest chromosomes (5, 21), this pattern was consistent and was particularly pronounced in the larger ones. For instance, the average recombination rate in the first and last 5 Mb of the largest chromosome (4) was 7.8 and 6.8 cM/Mb, whereas the segment ranging from 10 to 25 Mb exhibited an *c.* 20-fold lower rate (0.4 cM/Mb). The general pattern of periphery-biased recombination proved essentially insensitive to centromere position (e.g. compare chromosomes 7 and 8 in Fig. 1). Our data also suggested a tendency for the recombination rate to drop again right at the chromosome ends (e.g. chromosomes 1, 2, 4, 17 in Fig. 1). Formally testing this observation, however, would have required higher-resolution data, sampling the terminal domains more densely.

Comparing mean crossover number per meiosis among chromosomes revealed a lower limit of approximately one crossover for the chromosomes at the lower end of the size range (around 15 Mb) (Fig. 2A). With increasing chromosome length, the crossover number also increased ($r = 0.92$, permutation $P = 0.0001$), with the largest chromosomes (around 30 Mb) displaying *c.* 1.5 crossovers per meiosis. We also found a positive

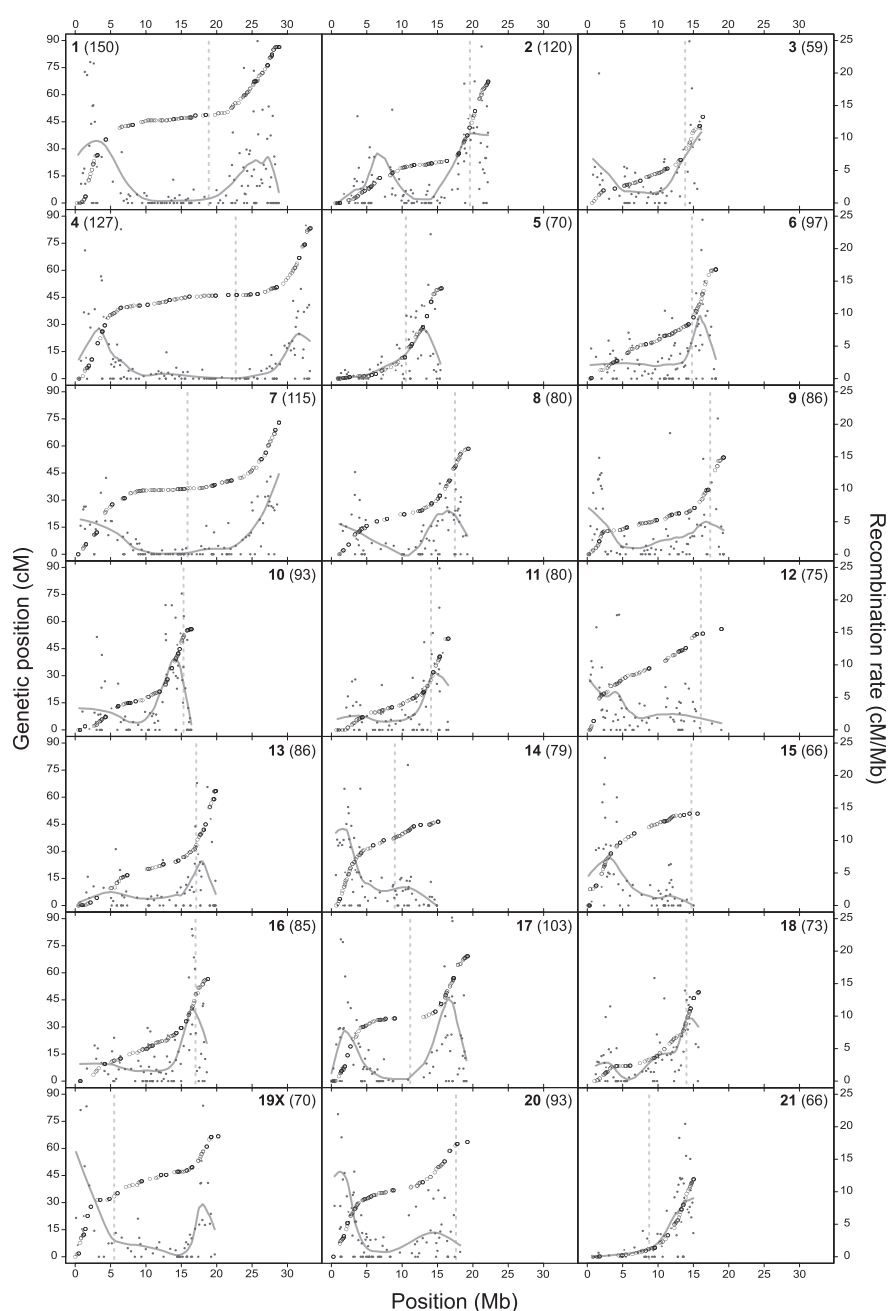


Fig. 1 Recombination along the 20 threespine stickleback autosomes (based on 1872 total markers and 282 F_2 individuals) and along the X chromosome (chromosome 19; based on 142 females). Marker number per chromosome is given in parentheses. The open circles (referring to the left axis) indicate the genetic map position of the markers in Kosambi centimorgan, plotted against their physical position in the genome (in megabases). The smaller grey dots (right axis) represent the average recombination rate in cM/Mb for the intervals defined by pairs of adjacent markers, plotted against the intervals' physical midpoint. The grey curves show the latter data smoothed by LOESS, with a polynomial degree of one and the smoothing span decreasing from 0.33 to 0.149 from the smallest to the largest chromosome to ensure a constant smoothing resolution across the panels. Dashed vertical lines specify centromere positions. Note the striking trend towards elevated recombination rate in the peripheral chromosome regions.

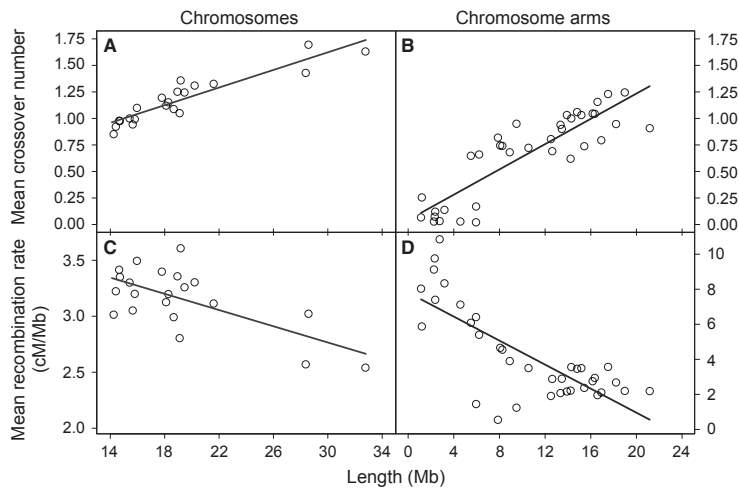


Fig. 2 Stickleback chromosomes display an approximate minimal crossover number of one per meiosis, and crossover number is related positively to chromosome length (A). A similar positive relationship exists between crossover number and chromosome *arm* length (B). Because the increase in crossover number with increasing chromosome length is not proportional, longer chromosomes display a lower average recombination rate (C). The same holds for chromosome arms (D). Note that (C) and (D) have different scales on the Y-axis.

association between chromosome *arm* length and crossover number ($r = 0.87$, $P = 0.0001$) (Fig. 2B). Along the short arms of telocentric and acrocentric chromosomes, crossovers occurred rarely. These relationships caused the average recombination rate to be higher on short chromosomes and chromosome arms than on longer ones (Fig. 2C, 2D; chromosomes: $r = -0.66$, $P = 0.0026$; arms: $r = -0.76$, $P = 0.0001$).

We found no indication of an overall difference in recombination rate between the sexes ($P = 0.982$); total autosomal map length was almost identical (1190 cM) for males and females. Analysing each chromosome separately also revealed only trivial sex-related differences in recombination rate (none of them remained significant when correcting for multiple testing). We further detected no significant QTL driving overall recombination rate on any of the 21 chromosomes (maximum LOD = 2.98; $P = 0.261$), keeping in mind the methodological limitations mentioned above.

Recombination and degeneration along the sex chromosome

In our F_2 population, XY recombination never occurred beyond the marker located at 1.75 Mb (Fig. 3). We thus demonstrate the presence of a small pseudoautosomal region (PAR), spanning *c.* 10% of the entire X chromosome only. The comparison of female with male RAD sequence coverage along the sex chromosome clearly revealed Y degeneration outside the PAR (Fig. 3). Moreover, the extent of degeneration was not uniform outside the PAR: within the segment from *c.* 12 Mb to the chromosome end opposed to the PAR, Y sequences generally did not align to the X reference. By contrast, the segment ranging from *c.* 2 to 12 Mb showed weaker

degeneration. Despite small sample size and hence more random noise, analogous analyses in the two natural populations produced very similar results supporting identical conclusions (Appendix S5, Supporting information).

Genetic divergence, genetic diversity and GC content

The prediction of a negative genome-wide association between recombination rate and F_{ST} was clearly confirmed (Table 1; Fig. 4A). Shifts in allele frequencies between populations were thus greater in low-recombination regions. As expected, this effect was stronger in the Boot lake–stream pair showing greater overall divergence (and hence higher variance in F_{ST}) than the Robert's pair (see Roesti *et al.* 2012a). The two complementary analyses of genetic diversity within populations also agreed with the prediction: all correlations were positive and generally highly significant (Table 1; Fig. 4B). Genetic diversity was thus reduced in marker intervals exhibiting a relatively low recombination rate. Finally, we found a striking positive broad-scale association between recombination rate and GC content across the genome (Fig. 5; $P = 0.0001$). Marker intervals showing relatively high recombination rates (around 10 cM/Mb or greater) displayed an *c.* 10% higher average GC content than intervals at the lower end of recombination rates (around 1 cM/Mb or lower).

Discussion

A major finding of our analysis of recombination in the threespine stickleback genome is the strong bias of crossover to occur primarily in the chromosome peripheries. This pattern confirms preliminary evidence from

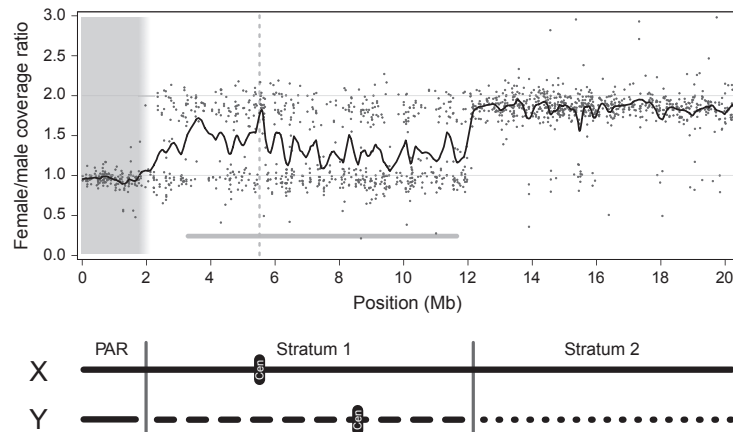


Fig. 3 Patterns of recombination and divergence between the X and Y chromosome. In the top panel, the abscissa gives the physical position along the reference X chromosome. The centromere is indicated by the dashed grey vertical line. The domain on the left shaded in grey indicates the extent of the pseudoautosomal region (PAR) where the gametologs still recombined in our cross (the PAR boundary lies between 1.75 and 2.22 Mb). The dots show the ratio of female to male sequence coverage across 100 individuals per sex for 1556 RAD loci (the black curve shows these data smoothed; degree = 1, span = 0.025). Within the PAR, the coverage ratio approximates unity (lower grey horizontal line), as expected for a DNA segment homochromatic between X and Y. Outside the PAR, many RAD loci display twofold higher sequence coverage in females than males (upper grey horizontal line), consistent with strong degeneration or loss of the X sequence on the Y. Note that two levels of Y degeneration ('evolutionary strata') are indicated (abutting at 12 Mb), the left one coinciding with the minimal size of a pericentric inversion on the Y inferred by Ross & Peichel (2008; visualized as heavy grey horizontal bar). On the bottom, we present the patterns of XY divergence inferred from our data in schematic form. Highly consistent patterns were also found when analysing natural population samples from Europe and Canada (see Appendix S5, Supporting information).

Table 1 Genome-wide associations between recombination rate and genetic population divergence, and between recombination rate and within-population genetic diversity. Divergence was quantified as F_{ST} between the lake and stream sample within the Boot and Robert's population pair. Genetic diversity within each of the four populations was expressed as both haplotype diversity (capturing allele frequency shifts) and the density of single nucleotide polymorphisms (SNPs) per RAD locus. All these metrics were averaged within the physical intervals defined by adjacent markers in the SNP panel used for genetic mapping ($N = 1783$ intervals)

	Population(s)	r	p
Genetic divergence (F_{ST})	Boot lake—stream	-0.2699	0.0001
	Robert's lake—stream	-0.1127	0.0001
Haplotype diversity	Boot lake	0.1184	0.0001
	Boot stream	0.0925	0.0022
	Robert's lake	0.0400	0.0929
	Robert's stream	0.0665	0.0113
SNP density	Boot lake	0.1873	0.0001
	Boot stream	0.1810	0.0001
	Robert's lake	0.2593	0.0001
	Robert's stream	0.1882	0.0001

low-resolution analyses in the species (Hohenlohe *et al.* 2012; Roesti *et al.* 2012a) and is consistent with results from genome-wide recombination studies in other

vertebrates (Borodin *et al.* 2008; Chowdhury *et al.* 2009; Backström *et al.* 2010; Wong *et al.* 2010; Bradley *et al.* 2011; Auton *et al.* 2012; Sandor *et al.* 2012; Tortereau *et al.* 2012), invertebrates (Rockman & Kruglyak 2009; Niehuis *et al.* 2010), plants (Akhunov *et al.* 2003; Anderson *et al.* 2003; but see Salomé *et al.* 2012) and yeast (Barton *et al.* 2008). This striking consistency across taxa implies a common mechanistic basis: crossovers seem to be initiated from the peripheries. Indeed, peripheral clustering of chromosomes during the meiotic prophase I is believed to play a key role in proper homolog pairing and probably also in crossover initiation (Scherthan *et al.* 1996; Roeder 1997; Harper *et al.* 2004; Brown *et al.* 2005; Naranjo & Corredor 2008). Peripheral crossover might also favour proper homolog dissociation (Colombo & Jones 1997; Hassold & Hunt 2001). Whatever the exact cause, the observed periphery bias in the distribution of crossovers in the stickleback genome (and many other genomes) implies strong mechanistic constraints on the distribution of recombination. Therefore, genetic information is reshuffled much more effectively in some genomic regions than in others.

Moreover, taking into account a slight underestimation of recombination (see Materials and methods), our data indicate that stickleback chromosomes display at least one crossover per meiosis. This is consistent with

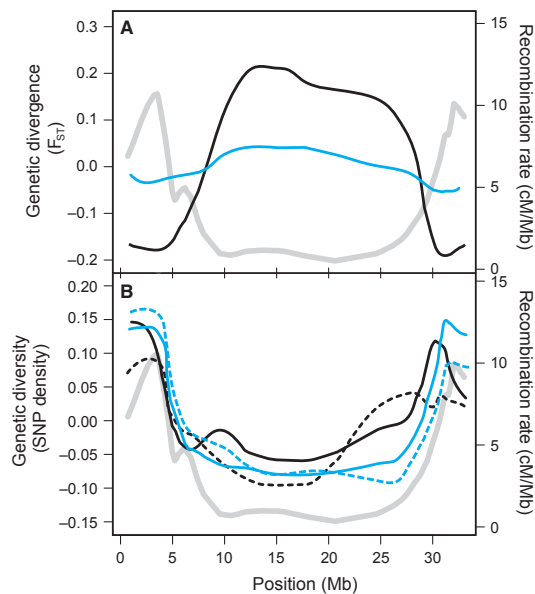


Fig. 4 Genetic divergence and genetic diversity in relation to recombination rate (shown as heavy grey line referring to the right axis in both panels) in natural lake and stream stickleback populations, exemplified for the largest chromosome (4). In (A), we show the smoothed (degree = 0, span = 0.35) magnitude of lake–stream divergence (F_{ST}) for the Boot (black) and Robert's (blue) population pair (for the sake of clarity, the underlying raw data points are not shown). To facilitate comparison, the data were centred to a mean of zero before smoothing. Note that divergence is greatest in the chromosome centre where recombination rate is lowest, an effect more pronounced in the Boot population pair showing much stronger overall divergence. In (B), we display smoothed genetic diversity, quantified as single nucleotide polymorphism density, for the lake (solid line) and stream (dashed line) population in the Boot (black) and Robert's (blue) population pair (data also centred). Note the strong and consistent positive association between genetic diversity and recombination rate.

the notion that one crossover per chromosome and meiosis is generally required for proper homolog segregation (Mather 1938; Baker *et al.* 1976; Roeder 1997; Smith & Nicolas 1998; Hassold & Hunt 2001), and reflects another mechanistic constraint on recombination. The widely accepted idea of one obligate crossover per chromosome *arm*, however, is not supported by our data (see also Borodin *et al.* 2008; Fledel-Alon *et al.* 2009): on acrocentric and telocentric stickleback chromosomes, the shorter arm rarely crosses over. We further find that the number of crossovers beyond one is a function of chromosome length. Standardized by their length, however, large chromosomes still exhibit a lower recombination rate than small chromosomes, the same also being true for chromosome arms.

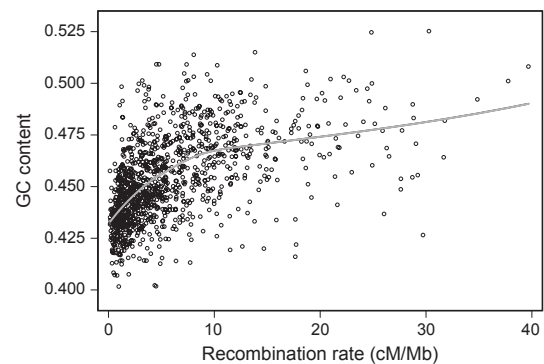


Fig. 5 Genome-wide relationship between GC content and recombination rate. The data points are the intervals defined by adjacent markers in the single nucleotide polymorphism panel used for genetic mapping ($N = 1783$ intervals). GC content was calculated for each marker interval by using information from the stickleback reference genome. The grey line shows the smoothed data (degree = 2, span = 0.85; a standard correlation produces $r = 0.49$).

Sex chromosome evolution

Sex chromosomes are generally thought to evolve from an ordinary pair of homologous autosomes that partly stop crossing over to prevent alleles at loci with sexually antagonistic effects from recombining (Bull 1983; Charlesworth & Charlesworth 2000; Charlesworth *et al.* 2005; Wilson & Makova 2009). This cessation of recombination should initiate the differentiation of the gametologs. While early karyotypic investigations in threespine stickleback found no evidence of heteromorphic sex chromosomes (Chen & Reisman 1970; Cuñado *et al.* 2002), recent investigations have indicated reduced recombination, chromosomal rearrangements and sequence divergence between the X and Y (Peichel *et al.* 2004; Ross & Peichel 2008; Shikano *et al.* 2011; Natri *et al.* 2013). These observations, based on a small number of markers, are greatly refined and extended by our sex chromosome analysis. We confirm that XY recombination is restricted to a small PAR, as suggested by Ross & Peichel (2008). The requirement of at least one crossover per meiosis thus implies a very high average recombination rate (*c.* 25 cM/Mb) across the PAR in males. This agrees with the estimation by Peichel *et al.* (2004) of a much greater distance between markers lying within the PAR in male than in female genetic maps (e.g. the genetic distance in the Paxton cross between the microsatellites Stn303 and Stn186, located at 0.4 and 1.9 Mb, is 27.3 cM in females and 47.7 cM in males).

A consequence of the cessation of recombination along most of the sex chromosome is that the region on the Y outside the PAR occurs in permanently

heterozygous state and at lower population size than the X. Both conditions are predicted to make selection on the Y less effective and hence to promote its degeneration (Felsenstein 1974; Charlesworth & Charlesworth 2000; Otto & Lenormand 2002; Charlesworth *et al.* 2005; Wilson & Makova 2009). Our results strongly support this view: outside the PAR, RAD loci often display only half the sequence coverage in males relative to females, consistent with substantial sequence degeneration (or loss) on the Y. Interestingly, our analysis further indicates two discrete levels of Y degeneration, with much stronger degeneration along the *c.* 8 Mb towards the chromosome end opposed to the PAR than along the *c.* 10 Mb adjacent to the PAR. Such 'evolutionary strata' (Lahn & Page 1999) have been found in mammals (Lahn & Page 1999; Sandstedt & Tucker 2004; Pearks Wilkerson *et al.* 2008), birds (Lawson Handley *et al.* 2004; Nam & Ellegren 2008) and plants (Bergero *et al.* 2007; Wang *et al.* 2012). To our knowledge, we here provide the first evidence for evolutionary strata in fish.

Evolutionary strata are generally taken as evidence that XY recombination ceased simultaneously across large domains of the evolving sex chromosome. An obvious way how this may happen is through chromosomal inversion. Indeed, a recent study using *in situ* fluorescent hybridization argued for a large pericentric inversion on the Y relative to the X, with breakpoints at *c.* 3 and 12 Mb (Ross & Peichel 2008). The evolutionary stratum adjacent to the PAR identified in our work matches this inversion almost perfectly and allows us to refine its physical boundaries. Threespine stickleback thus reinforce the view that recombination suppression along evolving sex chromosomes will primarily occur through inversion rather than crossover rate modifier genes (Charlesworth *et al.* 2005).

It would now be interesting to date the two bouts of recombination suppression underlying the evolutionary strata in the species based on sequence divergence between homologous loci on the X and Y (Lahn & Page 1999; Lawson Handley *et al.* 2004; Nam & Ellegren 2008; Pearks Wilkerson *et al.* 2008; Wang *et al.* 2012). We note that this might be difficult for stratum 2 if the remarkably strong degeneration detected in our analysis actually reflects deletion (Peichel *et al.* 2004). Clearly, however, patterns of XY divergence were already established prior to the split into Pacific and Atlantic stickleback clades (Appendix S5, Supporting information).

Consequences of heterogeneous recombination rate on genome evolution

The rate of recombination within a genomic region determines to which extent selection on a locus influences allele frequencies at neutral loci, and interferes

with selection on other loci, in its physical neighbourhood (Hill & Robertson 1966; Maynard Smith & Haigh 1974; Barton & Bengtsson 1986; Kaplan *et al.* 1989; Begun & Aquadro 1992; Nordborg *et al.* 1996; Charlesworth *et al.* 1997; Charlesworth 1998; Nachman 2002; Feder & Nosil 2010). Several types of selection (divergent, positive and background) should therefore increase divergence among populations and reduce genetic diversity within populations in low-recombination genomic regions relative to regions where recombination rate is higher.

Consistent with these predictions, we have recently shown that the magnitude of divergence between neighbouring lake and stream stickleback populations is dramatically biased towards chromosome centres (Roesti *et al.* 2012a). (Note that divergence in these young populations essentially reflects differential sorting of standing variation rather than novel mutations.) Using robust recombination rate data, we here demonstrate that elevated divergence in these population pairs is related to reduced recombination. Because lake and stream stickleback occupy selectively distinct environments (Berner *et al.* 2008, 2009), the divergence–recombination association almost certainly arises from within-chromosome variation in hitchhiking and/or introgression.

The present study further demonstrates reduced within-population genetic diversity in the chromosome centres relative to the peripheries, resulting in a genome-wide positive correlation between diversity and recombination rate. A similar correlation has previously been reported in a broad range of organisms (Begun & Aquadro 1992; Kraft *et al.* 1998; Nachman 2001; Tenaillon *et al.* 2001; Takahashi *et al.* 2004; Roselius *et al.* 2005; McGaugh *et al.* 2012). Given that a positive correlation between recombination rate and genetic diversity may also arise if recombination is directly mutagenic (Spencer *et al.* 2006; Webster & Hurst 2012; but see McGaugh *et al.* 2012), caution is generally warranted when inferring from the above correlation that recombination rate modulates the influence of selection on linked sites across the genome. In our lake–stream stickleback systems, however, the colocalization of elevated population divergence and reduced genetic diversity within young populations residing in selectively distinct environments provides clear support for such an indirect influence of recombination on genome evolution (see also Stoelting *et al.* 2013). The precise selective processes driving these patterns, however, remain to be elucidated.

In addition to these indirect (selective) effects, our study perhaps also points to a direct effect of recombination on stickleback genome evolution: large-scale bias in nucleotide composition. Across the genome, GC content is higher in regions displaying relatively elevated

recombination rate—that is, in the chromosome peripheries. Interestingly, the positive association between GC content and recombination rate seems as widespread as periphery bias in recombination rate; it has been reported in mammals (Jensen-Seaman *et al.* 2004; Spencer *et al.* 2006; Duret & Arndt 2008; Auton *et al.* 2012; Tortereau *et al.* 2012), birds (ICGSC 2004; Backström *et al.* 2010), insects (Niehuis *et al.* 2010; Stevison & Noor 2010; but see Comeron *et al.* 2012), plants (Muyle *et al.* 2011) and yeast (Gerton *et al.* 2000; Birdsell 2002). As hypothesized in other organisms, elevated GC content in the stickleback genome might represent a direct causal consequence of elevated recombination rate, given evidence of GC bias in the machinery correcting nucleotide mismatch in heteroduplex DNA formed during crossover initiation (GC-biased gene conversion; Brown & Jiricny 1987; Bill *et al.* 1998; Galtier *et al.* 2001; Birdsell 2002; Meunier & Duret 2004; Mancera *et al.* 2008; Duret & Galtier 2009; Muyle *et al.* 2011). Our correlational data, however, cannot address this causal hypothesis conclusively; direct experimental evidence is needed.

Methodological implications

In addition to the above influences on genome evolution, heterogeneous recombination rate within the genome has important methodological implications. Marker-based genome scans searching for signatures of divergent selection in the form of locally elevated divergence between ecologically distinct populations (Lewontin & Krakauer 1973; Beaumont & Nichols 1996; Luikart *et al.* 2003; Beaumont 2005; Nielsen 2005; Storz 2005) are becoming commonplace. What is generally ignored is that the distortion between physical and genetic maps will dilute the link between the selection coefficient on a locus and the magnitude of hitchhiking produced in its neutral neighbourhood (Roesti *et al.* 2012a; this study). In other words, a locus under selection is more likely to be detected when located in a low-recombination region where hitchhiking is more extensive. This bias should increase with decreasing marker resolution and with increasing sliding window size. The generality of chromosome periphery-biased recombination rate across taxa therefore raises a potential caveat to the interpretation of differentiation outliers in genome scans when combined physical and genetic map information is missing (i.e. 'anonymous' approaches; for one strategy to alleviate this difficulty when a physical map is available, see Roesti *et al.* 2012a). An analogous issue arises when interpreting the number and effect size of mapped QTL: within low-recombination regions, multiple loci of small effect are more likely to emerge as a single large-effect locus (Noor *et al.* 2001).

Finally, our study highlights the need for a reliably assembled genome for investigations of recombination and linkage. Assembly errors will inflate the genome-wide average crossover frequency, distort the recombination landscape and bias analyses of linkage disequilibrium along chromosomes. For instance, we find that a high-recombination island on chromosome 4 inferred in Hohenlohe *et al.* (2012; Fig. 2b in that study) coincides with the boundary of a scaffold anchored in the wrong sense within that chromosome (see our Appendix S1, Supporting information) and hence represents an artefact. The same assembly error also mimics long-distance linkage disequilibrium along this chromosome (Fig. 4a in Hohenlohe *et al.* 2012).

To summarize, our analysis of recombination in three-spine stickleback indicates strong constraints on the frequency and location of crossovers imposed by the mechanistic requirements of meiosis. At the same time, we demonstrate that recombination influences the genome profoundly, both by modulating the consequences of selection across the genome and perhaps by directly influencing nucleotide composition. We anticipate that our characterization of the recombination landscape will facilitate interpretations of genome scans and QTL mapping in the species, promote further investigations on sex chromosome evolution and pave the way for more detailed investigations of the determinants and consequences of recombination.

Acknowledgements

The following contributions made this study possible and are most gratefully acknowledged: F. Hofmann (SFFN—Inspection de la pêche VD) and R. Kistler (fisheries authorities of the canton Thurgau) provided sampling permits for the cross populations, and A.-C. Grandchamp aided sampling. A. Hendry supported sampling of the field populations financially. B. Cresko and C. Peichel provided input on the cross and marker generation design at an early stage. B. Egger, H. Gante, A. Indermaur, W. Salzburger, A. Theys and P. Vonlanthen helped rear the mapping cross. W. Salzburger generously shared laboratory resources and infrastructure. B. Aeschbach and N. Boileau facilitated wet laboratory work. Illumina sequencing was performed by C. Beisel and I. Nissen at the Quantitative Genomics Facility, D-BSSE, ETH Zürich. J. Urton shared information on centromere positions, and M. Noor and two other reviewers provided valuable suggestions to improve the manuscript. MR was supported by a Swiss National Science Foundation (SNF) Sinergia grant (CRSII3_136293) to M. Sanchez, H. Furrer, and W. Salzburger. DB was supported by a SNF Ambizione fellowship (PZ00P3_126391/1) and by the Research Fund of the University of Basel.

References

Akhunov ED, Goodyear AW, Geng S *et al.* (2003) The organization and rate of evolution of wheat genomes are correlated

- with recombination rates along chromosome arms. *Genome Research*, **13**, 753–763.
- Albert AYK, Sawaya S, Vines TH *et al.* (2008) The genetics of adaptive shape shift in stickleback: pleiotropy and effect size. *Evolution*, **62**, 76–85.
- Anderson LK, Doyle GG, Brigham B *et al.* (2003) High-resolution crossover maps for each bivalent of *Zea mays* using recombination nodules. *Genetics*, **165**, 849–865.
- Auton A, Fledel-Alon A, Pfeifer S *et al.* (2012) A fine-scale chimpanzee genetic map from population sequencing. *Science*, **336**, 193–198.
- Backström N, Forstmeier W, Schielzeth H *et al.* (2010) The recombination landscape of the zebra finch *Taeniopygia guttata* genome. *Genome Research*, **20**, 485–495.
- Baird NA, Etter PD, Atwood TS, *et al.* (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE*, **3**, e3376.
- Baker BS, Carpenter ATC, Esposito MS, Esposito RE, Sandler L (1976) The genetic control of meiosis. *Annual Review of Genetics*, **10**, 53–134.
- Barton N, Bengtsson BO (1986) The barrier to genetic exchange between hybridizing populations. *Heredity*, **57**, 357–376.
- Barton AB, Peksos MR, Kurvathi RS, Kaback DB (2008) Meiotic recombination at the ends of chromosomes in *Saccharomyces cerevisiae*. *Genetics*, **179**, 1221–1235.
- Beaumont MA (2005) Adaptation and speciation: what can F_{st} tell us? *Trends in Ecology & Evolution*, **20**, 435–440.
- Beaumont MA, Nichols RA (1996) Evaluating loci for use in the genetic analysis of population structure. *Proceedings of the Royal Society of London, Series B: Biological Sciences*, **263**, 1619–1626.
- Begun DJ, Aquadro CF (1992) Levels of naturally occurring DNA polymorphism correlate with recombination rates in *Drosophila melanogaster*. *Nature*, **356**, 519–520.
- Bergero R, Forrest A, Kamau E, Charlesworth D (2007) Evolutionary strata on the X chromosomes of the dioecious plant *Silene latifolia*: evidence from new sex-linked genes. *Genetics*, **175**, 1945–1954.
- Berner D, Adams DC, Grandchamp A-C, Hendry AP (2008) Natural selection drives patterns of lake-stream divergence in stickleback foraging morphology. *Journal of Evolutionary Biology*, **21**, 1653–1665.
- Berner D, Grandchamp A-C, Hendry AP (2009) Variable progress toward ecological speciation in parapatry: stickleback across eight lake-stream transitions. *Evolution*, **63**, 1740–1753.
- Berner D, Roesti M, Hendry AP, Salzburger W (2010) Constraints on speciation suggested by comparing lake-stream stickleback divergence across two continents. *Molecular Ecology*, **19**, 4963–4978.
- Bill CA, Duran WA, Miselis NR, Nickoloff JA (1998) Efficient repair of all types of single-base mismatches in recombination intermediates in Chinese hamster ovary cells: competition between long-patch and G-T glycosylase-mediated repair of G-T mismatches. *Genetics*, **149**, 1935–1943.
- Birdsell JA (2002) Integrating genomics, bioinformatics, and classical genetics to study the effects of recombination on genome evolution. *Molecular Biology and Evolution*, **19**, 1181–1197.
- Borodin PM, Karamysheva TV, Belonogova NM *et al.* (2008) Recombination map of the common shrew, *Sorex araneus* (eulipotyphla, mammalia). *Genetics*, **178**, 621–632.
- Bradley KM, Breyer JP, Melville DB, Broman KW, Knapik EW, Smith JR (2011) An SNP-based linkage map for zebrafish reveals sex determination loci. *G3: Genes, Genomes, Genetics*, **1**, 3–9.
- Broman KW, Sen S (2009) *A Guide to QTL Mapping With R/qtl*. Springer, New York City, New York.
- Brown TC, Jiricny J (1987) A specific mismatch repair event protects mammalian cells from loss of 5-methylcytosine. *Cell*, **50**, 945–950.
- Brown PW, Judis LA, Chan ER *et al.* (2005) Meiotic synapsis proceeds from a limited number of subtelomeric sites in the human male. *American Journal of Human Genetics*, **77**, 556–566.
- Bull JJ (1983) *Evolution of Sex Determining Mechanisms*. Benjamin Cummings, Menlo Park, CA.
- Burt A (2000) Sex, recombination, and the efficacy of selection – was Weismann right? *Evolution*, **54**, 337–351.
- Charlesworth B (1998) Measures of divergence between populations and the effect of forces that reduce variability. *Molecular Biology and Evolution*, **15**, 538–543.
- Charlesworth B, Charlesworth D (2000) The degeneration of Y chromosomes. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, **355**, 1563–1572.
- Charlesworth B, Nordborg M, Charlesworth D (1997) The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations. *Genetical Research*, **70**, 155–174.
- Charlesworth D, Charlesworth B, Marais G (2005) Steps in the evolution of heteromorphic sex chromosomes. *Heredity*, **95**, 118–128.
- Chen TR, Reisman HM (1970) Comparative chromosome study of North American species of sticklebacks (Teleostei – Gasterosteidae). *Cytogenetics*, **9**, 321–332.
- Chowdhury R, Bois PRJ, Feingold E, Sherman SL, Cheung VG (2009) Genetic analysis of variation in human meiotic recombination. *PLoS Genetics*, **5**, e1000648.
- Cleveland WS (1979) Robust locally weighted regression and smoothing scatterplots. *Journal of American Statistical Association*, **74**, 829–836.
- Colombo PC, Jones GH (1997) Chiasma interference is blind to centromeres. *Heredity*, **79**, 214–227.
- Comeron JM, Ratnappan R, Bailin S (2012) The many landscapes of recombination in *Drosophila melanogaster*. *PLoS Genetics*, **8**, e1002905.
- Cuñado N, Barrios J, San Miguel E *et al.* (2002) Synaptonemal complex analysis in oocytes and spermatocytes of threespine stickleback *Gasterosteus aculeatus* (Teleostei, Gasterosteidae). *Genetica*, **114**, 53–56.
- Duret L, Arndt PF (2008) The impact of recombination on nucleotide substitutions in the human genome. *PLoS Genetics*, **4**, e1000071.
- Duret L, Galtier N (2009) Biased gene conversion and the evolution of mammalian genomic landscapes. *Annual Review of Genomics and Human Genetics*, **10**, 285–311.
- Feder JL, Nosil P (2010) The efficacy of divergence hitchhiking in generating genomic islands during ecological speciation. *Evolution*, **64**, 1729–1747.
- Felsenstein J (1974) The evolutionary advantage of recombination. *Genetics*, **78**, 737–756.
- Fledel-Alon A, Wilson DJ, Broman KW *et al.* (2009) Broad-scale recombination patterns underlying proper disjunction in humans. *PLoS Genetics*, **5**, e1000658.

- Galtier N, Piganeau G, Mouchiroud D, Duret L (2001) GC-content evolution in mammalian genomes: the biased gene conversion hypothesis. *Genetics*, **159**, 907–911.
- Gerton JL, DeRisi J, Shroff R *et al.* (2000) Global mapping of meiotic recombination hotspots and coldspots in the yeast *Saccharomyces cerevisiae*. *Proceedings of the National Academy of Sciences of the United States of America*, **97**, 11383–11390.
- Greenwood AK, Jones FC, Chan YF *et al.* (2011) The genetic basis of divergent pigment patterns in juvenile threespine sticklebacks. *Heredity*, **107**, 155–166.
- Harper L, Golubovskaya I, Cande WZ (2004) A bouquet of chromosomes. *Journal of Cell Science*, **117**, 4025–4032.
- Hassold T, Hunt P (2001) To err (meiotically) is human: the genesis of human aneuploidy. *Nature Reviews Genetics*, **2**, 280–291.
- Hill WG, Robertson A (1966) Effect of linkage on limits to artificial selection. *Genetical Research*, **8**, 269–294.
- Hohenlohe PA, Bassham S, Currey M, Cresko WA (2012) Extensive linkage disequilibrium and parallel adaptive divergence across threespine stickleback genomes. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, **367**, 395–408.
- ICGSC (2004) Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature*, **432**, 695–716.
- Jensen-Seaman MI, Furey TS, Payseur BA *et al.* (2004) Comparative recombination rates in the rat, mouse, and human genomes. *Genome Research*, **14**, 528–538.
- Jones FC, Grabherr MG, Chan YF *et al.* (2012) The genomic basis of adaptive evolution in three spine sticklebacks. *Nature*, **484**, 55–61.
- Kaplan NL, Hudson RR, Langley CH (1989) The hitchhiking effect revisited. *Genetics*, **123**, 887–899.
- Kraft T, Sall T, Magnusson-Rading I, Nilsson NO, Hallden C (1998) Positive correlation between recombination rates and levels of genetic variation in natural populations of sea beet (*Beta vulgaris* subsp. *maritima*). *Genetics*, **150**, 1239–1244.
- Lahn BT, Page DC (1999) Four evolutionary strata on the human X chromosome. *Science*, **286**, 964–967.
- Lawson Handley L-J, Cepitis H, Ellegren H (2004) Evolutionary strata on the chicken Z chromosome: implications for sex chromosome evolution. *Genetics*, **167**, 367–376.
- Lewontin RC, Krakauer J (1973) Distribution of gene frequency as a test of theory of selective neutrality of polymorphisms. *Genetics*, **74**, 175–195.
- Li H, Handsaker B, Wysoker A *et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
- Luikart G, England PR, Tallmon D, Jordan S, Taberlet P (2003) The power and promise of population genomics: from genotyping to genome typing. *Nature Reviews Genetics*, **4**, 981–994.
- Mancera E, Bourgon R, Brozzi A, Huber W, Steinmetz LM (2008) High-resolution mapping of meiotic crossovers and non-crossovers in yeast. *Nature*, **454**, 479–485.
- Manly BFJ (2007) *Randomization, Bootstrap and Monte Carlo Methods in Biology*, 3rd edn. Chapman & Hall, Boca Raton.
- Mather K (1938) Crossing-over. *Biological Reviews*, **13**, 252–292.
- Maynard Smith J, Haigh J (1974) Hitch-hiking effect of a favorable gene. *Genetical Research*, **23**, 23–35.
- McCaugh SE, Heil CSS, Manzano-Winkler B *et al.* (2012) Recombination modulates how selection affects linked sites in *Drosophila*. *PLoS Biology*, **10**, e1001422.
- Meunier J, Duret L (2004) Recombination drives the evolution of GC-content in the human genome. *Molecular Biology and Evolution*, **21**, 984–990.
- Morgan M, Anders S, Lawrence M *et al.* (2009) ShortRead: a bioconductor package for input, quality assessment and exploration of high-throughput sequence data. *Bioinformatics*, **25**, 2607–2608.
- Moser D, Roesti M, Berner D (2012) Repeated lake-stream divergence in stickleback life history within a Central European lake basin. *PLoS ONE*, **7**, e50620.
- Muyle A, Serres-Giardi L, Ressayre A, Escobar J, Glemin S (2011) GC-biased gene conversion and selection affect GC content in the *Oryza* genus (rice). *Molecular Biology and Evolution*, **28**, 2695–2706.
- Nachman MW (2001) Single nucleotide polymorphisms and recombination rate in humans. *Trends in Genetics*, **17**, 481–485.
- Nachman MW (2002) Variation in recombination rate across the genome: evidence and implications. *Current Opinion in Genetics & Development*, **12**, 657–663.
- Nam K, Ellegren H (2008) The chicken (*Gallus gallus*) Z chromosome contains at least three nonlinear evolutionary strata. *Genetics*, **180**, 1131–1136.
- 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.
- Natri HM, Shikano T, Merilä J (2013) Progressive recombination suppression and differentiation in recently evolved neo-sex chromosomes. *Molecular Biology and Evolution*, in press.
- Nei M, Tajima F (1981) DNA polymorphism detectable by restriction endonucleases. *Genetics*, **97**, 145–163.
- Niehuis O, Gibson JD, Rosenberg MS *et al.* (2010) Recombination and its impact on the genome of the haplodiploid parasitoid wasp *Nasonia*. *PLoS ONE*, **5**, e8597.
- Nielsen R (2005) Molecular signatures of natural selection. *Annual Review of Genetics*, **39**, 197–218.
- Noor MAF, Cunningham AL, Larkin JC (2001) Consequences of recombination rate variation on quantitative trait locus mapping studies: simulations based on the *Drosophila melanogaster* genome. *Genetics*, **159**, 581–588.
- Nordborg M, Charlesworth B, Charlesworth D (1996) The effect of recombination on background selection. *Genetical Research*, **67**, 159–174.
- Otto SP, Barton NH (1997) The evolution of recombination: removing the limits to natural selection. *Genetics*, **147**, 879–906.
- Otto SP, Lenormand T (2002) Resolving the paradox of sex and recombination. *Nature Reviews Genetics*, **3**, 252–261.
- Pearks Wilkerson AJ, Raudsepp T, Graves T *et al.* (2008) Gene discovery and comparative analysis of X-degenerate genes from the domestic cat Y chromosome. *Genomics*, **92**, 329–338.
- Peichel CL, Ross JA, Matson CK *et al.* (2004) The master sex-determination locus in threespine sticklebacks is on a nascent Y chromosome. *Current Biology*, **14**, 1416–1424.
- R Development Core Team (2012) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Austria, Vienna.
- Rockman MV, Kruglyak L (2009) Recombinational landscape and population genomics of *Caenorhabditis elegans*. *PLoS Genetics*, **5**, e1000419.
- Roeder GS (1997) Meiotic chromosomes: it takes two to tango. *Genes & Development*, **11**, 2600–2621.

- Roesti M, Hendry AP, Salzburger W, Berner D (2012a) Genome divergence during evolutionary diversification as revealed in replicate lake-stream stickleback population pairs. *Molecular Ecology*, **21**, 2852–2862.
- Roesti M, Salzburger W, Berner D (2012b) Uninformative polymorphisms bias genome scans for signatures of selection. *BMC Evolutionary Biology*, **12**, 94.
- Roselius K, Stephan W, Stadler T (2005) The relationship of nucleotide polymorphism, recombination rate and selection in wild tomato species. *Genetics*, **171**, 753–763.
- Ross JA, Peichel CL (2008) Molecular cytogenetic evidence of rearrangements on the Y chromosome of the threespine stickleback fish. *Genetics*, **179**, 2173–2182.
- Salomé PA, Bomblies K, Fitz J *et al.* (2012) The recombination landscape in *Arabidopsis thaliana* F2 populations. *Heredity*, **108**, 447–455.
- Sandor C, Li W, Coppieters W *et al.* (2012) Genetic variants in REC8, RNF212, and PRDM9 influence male recombination in cattle. *PLoS Genetics*, **8**, e1002854.
- Sandstedt SA, Tucker PK (2004) Evolutionary strata on the mouse X chromosome correspond to strata on the human X chromosome. *Genome Research*, **14**, 267–272.
- Scherthan H, Weich S, Schwegler H *et al.* (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.
- Shikano T, Natri HM, Shimada Y, Merilä J (2011) High degree of sex chromosome differentiation in stickleback fishes. *BMC Genomics*, **12**, 474.
- Smith KN, Nicolas A (1998) Recombination at work for meiosis. *Current Opinion in Genetics & Development*, **8**, 200–211.
- Smukowski CS, Noor MAF (2011) Recombination rate variation in closely related species. *Heredity*, **107**, 496–508.
- Spencer CCA, Deloukas P, Hunt S *et al.* (2006) The influence of recombination on human genetic diversity. *PLoS Genetics*, **2**, 1375–1385.
- Stevison LS, Noor MAF (2010) Genetic and evolutionary correlates of fine-scale recombination rate variation in *Drosophila persimilis*. *Journal of Molecular Evolution*, **71**, 332–345.
- Stoelting KN, Nipper R, Lindtke D *et al.* (2013) Genomic scan for single nucleotide polymorphisms reveals patterns of divergence and gene flow between ecologically divergent species. *Molecular Ecology*, **22**, 842–855.
- Storz JF (2005) Using genome scans of DNA polymorphism to infer adaptive population divergence. *Molecular Ecology*, **14**, 671–688.
- Takahashi A, Liu YH, Saitou N (2004) Genetic variation versus recombination rate in a structured population of mice. *Molecular Biology and Evolution*, **21**, 404–409.
- Tenaillon MI, Sawkins MC, Long AD *et al.* (2001) Patterns of DNA sequence polymorphism along chromosome 1 of maize (*Zea mays* ssp. *mays* L.). *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 9161–9166.
- Tortoreau F, Servin B, Frantz L *et al.* (2012) A high density recombination map of the pig reveals a correlation between sex-specific recombination and GC content. *BMC Genomics*, **13**, 586.
- Urton JR, McCann SM, Peichel CL (2011) Karyotype differentiation between two stickleback species (Gasterosteidae). *Cytogenetic and Genome Research*, **135**, 150–159.
- Wang J, Naa J-K, Yu Q *et al.* (2012) Sequencing papaya X and Y chromosomes reveals molecular basis of incipient sex chromosome evolution. *Proceedings of the National Academy of Sciences of the United States of America*, **109**, 13710–13715.
- Webster MT, Hurst LD (2012) Direct and indirect consequences of meiotic recombination: implications for genome evolution. *Trends in Genetics*, **28**, 101–109.
- Wilson MA, Makova KD (2009) Genomic analyses of sex chromosome evolution. *Annual Review of Genomics and Human Genetics*, **10**, 333–354.
- Wong AK, Ruhe AL, Dumont BL *et al.* (2010) A comprehensive linkage map of the dog genome. *Genetics*, **184**, 595–605.

D.B. and M.R. designed the study and performed the analyses; D.B. produced the cross; M.R. and D.M. generated the R.A.D. libraries; D.B., M.R. and D.M. produced the marker data; D.B. wrote the manuscript, with input from M.R.

Data accessibility

Reassembled threespine stickleback reference genome: Dryad digital repository (doi:10.5061/dryad.846nj).

Genotype data used for final linkage map construction and the corresponding linkage map: online supporting information.

Supporting information

Additional supporting information may be found in the online version of this article.

Appendix S1 Schematic overview of the stickleback genome reassembly.

Appendix S2 Genotype data used for estimating genetic distances on the 20 autosomes.

Appendix S3 Genotype data used for estimating genetic distances on the sex chromosome (19X).

Appendix S4 Physical and genetic map positions for all 1872 markers.

Appendix S5 Patterns of Y chromosome degeneration revealed by sex-specific RAD locus coverage along the X chromosome in an Atlantic-derived and a Pacific-derived natural population.

Chapter 3

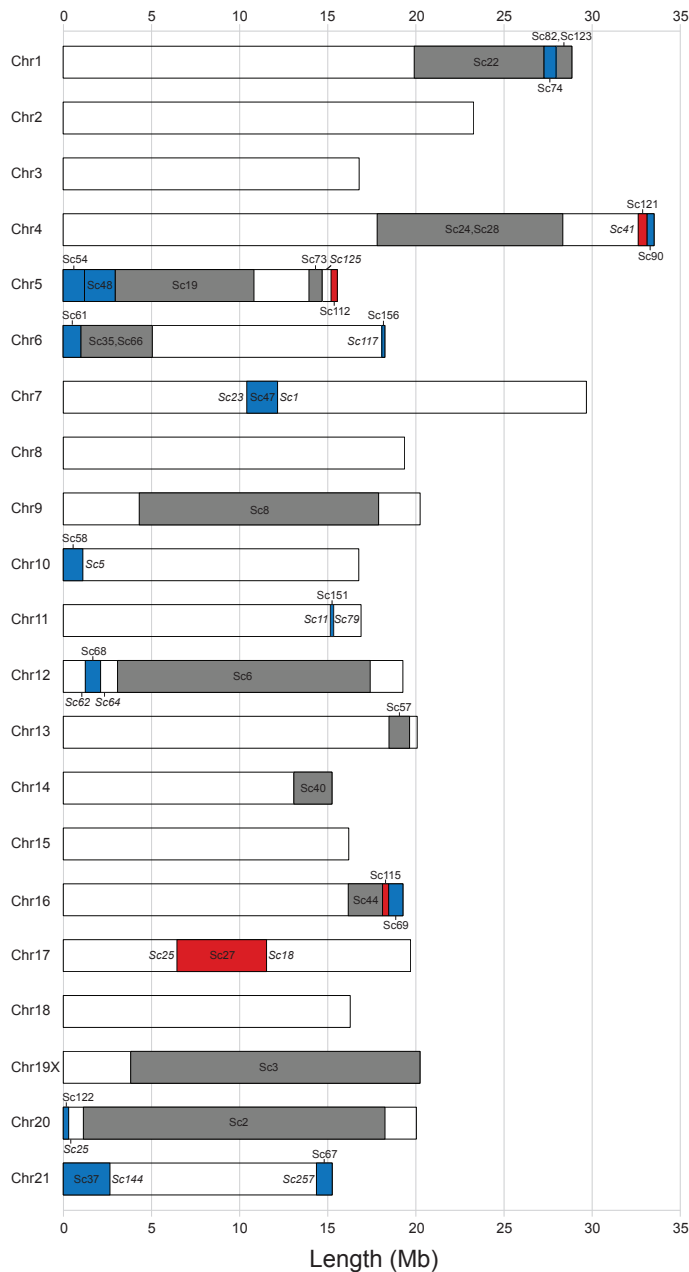
Recombination in the threespine stickleback
genome—patterns and consequences

Supplementary material

Appendices S2-S4 to Roesti et al. (2013) can be obtained online under:
<http://onlinelibrary.wiley.com/doi/10.1111/mec.12322/supinfo>

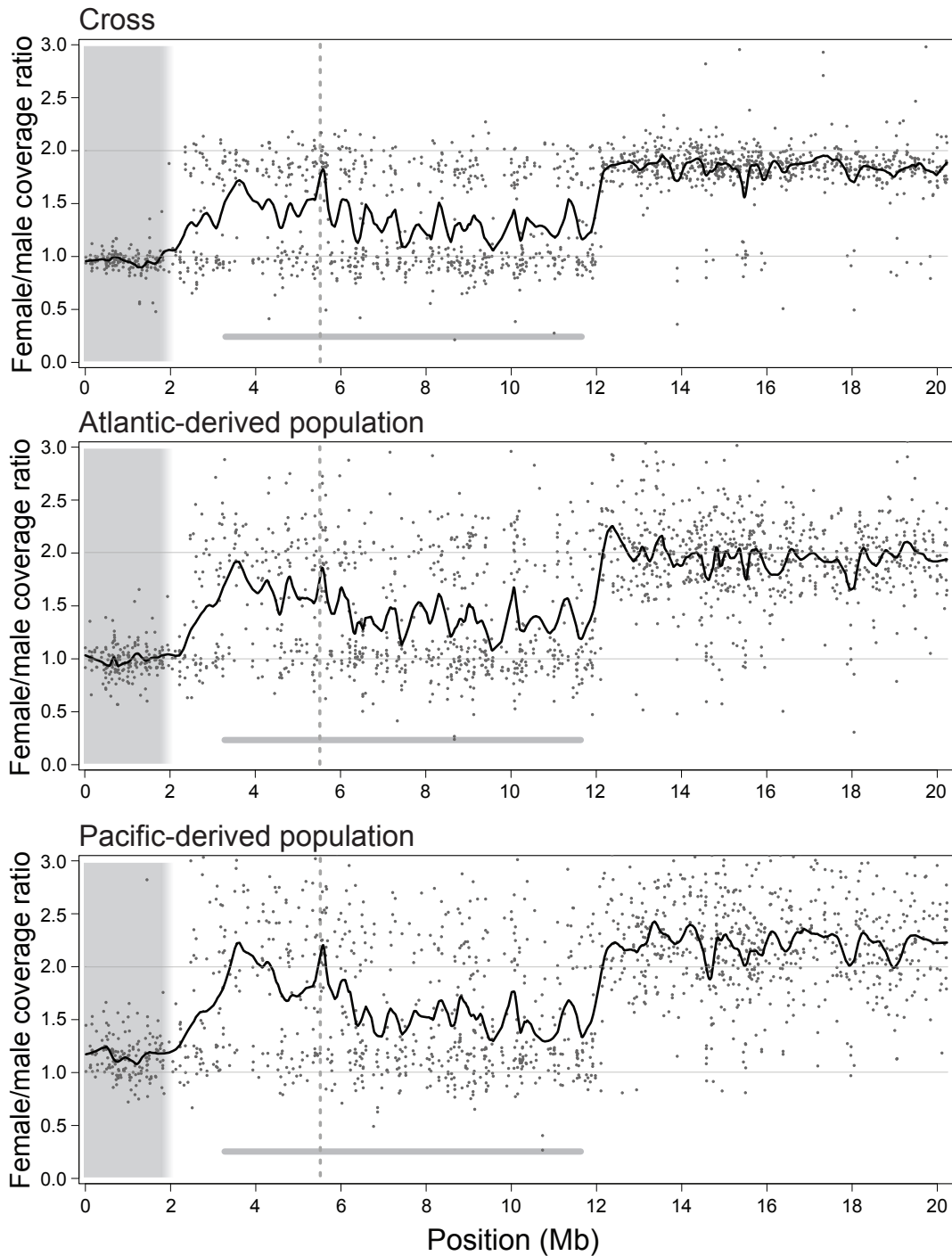
Appendix S1, Roesti et al. 2013

Reassembly of the Broad S1 threespine stickleback reference genome according to a de novo linkage map based on 282 F2 individuals and 2,165 genome-wide single nucleotide polymorphisms (SNPs). Each bar represents a chromosome (1 - 21); the length of the bar gives the total chromosome span (in megabase). Correctly assembled chromosome regions are shown in white. Scaffolds originally placed in the correct chromosome but in incorrect sense are shown in gray. Previously unanchored scaffolds that were integrated into the chromosomes are shown in blue (if incorporated in normal sense) and red (if incorporated in reverse sense). Each scaffold that was reversed and/or relocated is named (e.g., Sc22). For relocated scaffolds, we also provide the name of the flanking scaffold on each side. Note that scaffolds shorter than 140 kb were ignored. These scaffolds were generally represented by 1 - 3 SNPs only and typically mapped to chromosome ends.



Appendix S5, Roesti et al. 2013

Degeneration of the threespine stickleback Y chromosome revealed by female versus male RAD sequence coverage, compared across analyses based on different populations. The top panel is a copy of the graphic shown in Fig. 3, drawn based on data from our F2 population (100 individuals per sex). The middle and bottom panels were drawn by following the same plotting conventions, but are based on samples from two natural populations independently evolved in geographic isolation and from different marine ancestors (Atlantic, CHE; Pacific, Boot Lake). Sample size in these latter analyses was 13 -14 per sex and population. Note the striking consistency in female to male coverage across the different analyses.



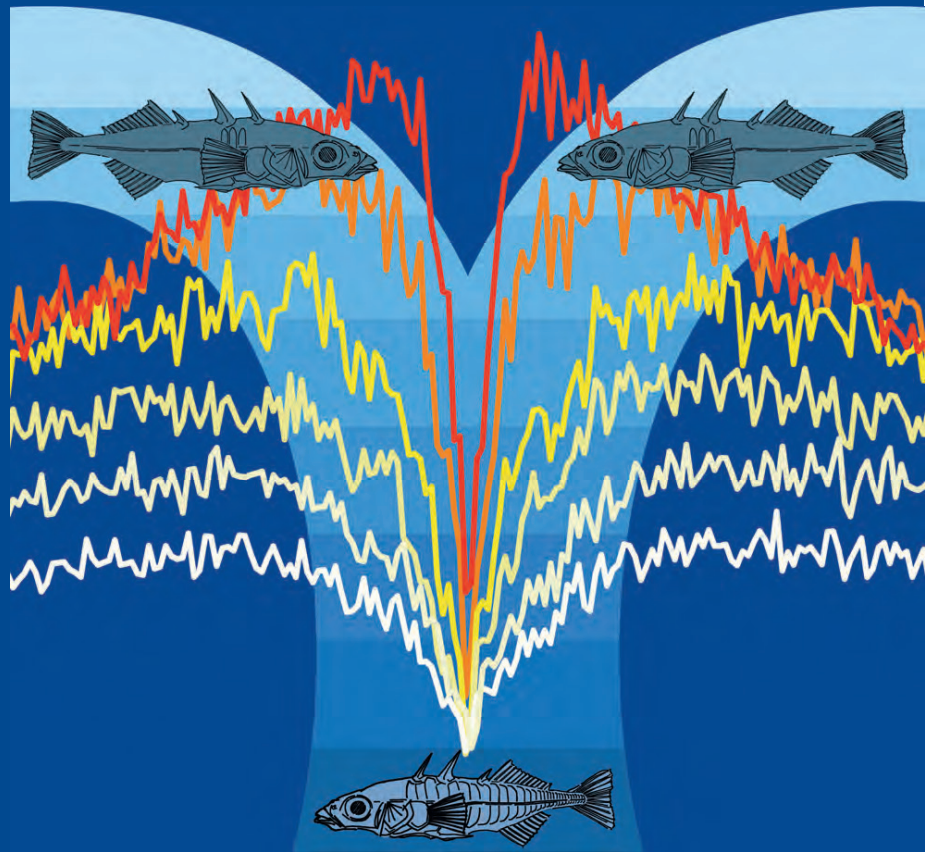
Chapter 4

The genomic signature of parallel adaptation from
shared genetic variation

VOLUME 23
NUMBER 16
AUGUST
2014

MOLECULAR ECOLOGY

FROM THE COVER: The genomic signature of parallel adaptation from shared genetic variation. See pp. 3944–3956.



Published by
WILEY Blackwell

FROM THE COVER

The genomic signature of parallel adaptation from shared genetic variation

MARIUS ROESTI,* SERGEY GAVRILETS,† ANDREW P. HENDRY,‡ WALTER SALZBURGER* and DANIEL BERNER*

*Zoological Institute, University of Basel, Vesalgasse 1, 4051 Basel, Switzerland, †Department of Ecology and Evolutionary Biology and Department of Mathematics, National Institute for Mathematical and Biological Synthesis (NIMBioS), University of Tennessee, Knoxville, TN 37996, USA, ‡Department of Biology and Redpath Museum, McGill University, 859 Sherbrooke St W., Montreal, Quebec, Canada

Abstract

Parallel adaptation is common and may often occur from shared genetic variation, but the genomic consequences of this process remain poorly understood. We first use individual-based simulations to demonstrate that comparisons between populations adapted in parallel to similar environments from shared variation reveal a characteristic genomic signature around a selected locus: a low-divergence valley centred at the locus and flanked by twin peaks of high divergence. This signature is initiated by the hitchhiking of haplotype tracts differing between derived populations in the broader neighbourhood of the selected locus (driving the high-divergence twin peaks) and shared haplotype tracts in the tight neighbourhood of the locus (driving the low-divergence valley). This initial hitchhiking signature is reinforced over time because the selected locus acts as a barrier to gene flow from the source to the derived populations, thus promoting divergence by drift in its close neighbourhood. We next empirically confirm the peak-valley-peak signature by combining targeted and RAD sequence data at three candidate adaptation genes in multiple marine (source) and freshwater (derived) populations of threespine stickleback. Finally, we use a genome-wide screen for the peak-valley-peak signature to discover additional genome regions involved in parallel marine-freshwater divergence. Our findings offer a new explanation for heterogeneous genomic divergence and thus challenge the standard view that peaks in population divergence harbour divergently selected loci and that low-divergence regions result from balancing selection or localized introgression. We anticipate that genome scans for peak-valley-peak divergence signatures will promote the discovery of adaptation genes in other organisms.

Keywords: barrier to gene flow, evolutionary genomics, *Gasterosteus aculeatus*, genetic hitchhiking, genome scan, population divergence

Received 21 January 2014; revision received 12 March 2014; accepted 12 March 2014

Introduction

Understanding how selection shapes the genome and identifying the loci underlying adaptive divergence are major goals of biology (Wu 2001; Nielsen 2005; Stinchcombe & Hoekstra 2008; Nosil & Schluter 2011; Feder *et al.* 2012). Recent studies have indicated that genomic

differentiation between diverging populations can be highly heterogeneous and can involve selection on numerous loci throughout the genome, with some of these loci now having been identified (e.g. Hohenlohe *et al.* 2010; Lawniczak *et al.* 2010; Fournier-Level *et al.* 2011; Jones *et al.* 2012b; Nadeau *et al.* 2012; Roesti *et al.* 2012a; Renaut *et al.* 2013; Streisfeld *et al.* 2013). Nevertheless, understanding how evolutionary processes cause heterogeneous genomic divergence remains challenging (e.g. Slatkin & Wiehe 1998; Barton 2000;

Correspondence: Daniel Berner, Fax: +41 (0) 61 267 0301; E-mail: daniel.berner@unibas.ch

Hermisson & Pennings 2005; Excoffier & Ray 2008; Bierne 2010; Feder & Nosil 2010; Bierne *et al.* 2011; Roesti *et al.* 2012a, 2013; reviewed in Wu 2001; Nosil *et al.* 2009). Traditional population genetic theory has primarily focused on a scenario in which a new genetic variant arises by mutation in a population colonizing a new environment (hereafter called a 'derived' population) where the variant is beneficial (Orr 1998; Barrett & Schluter 2008; Messer & Petrov 2013). The new genetic variant is then expected to fix in the derived population, whereas the initial genetic variant remains favoured and is thus retained in the 'source' population inhabiting the ancestral environment. Consequently, the source and derived populations are differentiated at the locus under divergent selection and, due to genetic hitchhiking, also in the selectively neutral genetic neighbourhood of that locus (Maynard Smith & Haigh 1974; Kaplan *et al.* 1989). Genomic regions of high population divergence, as identified in marker-based genome scans, are thus generally assumed to harbour genes involved in adaptive divergence (Nielsen 2005; Storz 2005).

This traditional theoretical scenario might not be adequate when adaptation occurs from standing (pre-existing) genetic variation rather than from novel mutations (Hermisson & Pennings 2005; Barrett & Schluter 2008; Pritchard *et al.* 2010). This realization has stimulated theory focusing on 'soft' selective sweeps, where a novel genetic variant is segregating in an ancestral source population before becoming selected in a derived population. In this case, the divergence signature driven by the selective sweep will be weakened relative to the classical 'hard' sweep expected from a novel mutation (Hermisson & Pennings 2005; Barrett & Schluter 2008; Messer & Petrov 2013). The reason is that the derived variant can become associated through recombination with diverse genetic backgrounds in the ancestral population before the derived population becomes established in the new environment. This diversity reduces the selective sweep in the locus' genomic neighbourhood when the variant eventually becomes selected. Although hard and soft sweep models differ in the age (or origin) of the selected variant and in the expected strength of the associated selective signature, they share the focus on comparing populations inhabiting *selectively different* environments (i.e. source vs. derived).

In this study, we consider the patterns of genomic divergence that might be expected among multiple *derived* populations adapting in parallel to *selectively similar* environments. We scrutinize these genomic patterns through theoretical modelling and through targeted and genome-wide sequencing in multiple natural populations of threespine stickleback fish (*Gasterosteus aculeatus*) that have adapted in parallel to freshwater

environments from a common marine source population. We find that a locus involved in parallel adaptation from shared genetic variation generates a novel and characteristic pattern of genomic divergence, which provides a new perspective on how to interpret high- and low-divergence outliers detected in genome scans.

Materials and methods

Models of parallel adaptation from shared genetic variation

We developed individual-based simulation models in which multiple derived populations diverge independently into a selectively novel environment from a shared source population inhabiting a selectively different, ancestral environment. A detailed description of these simulations is provided in the Methods S1 (Supporting information), so that we here give a brief overview only. Individuals are represented by a single haploid chromosome. The centre of the chromosome holds a locus under divergent selection between the ancestral and novel environment, with the allele favourable in the novel environment occurring at low frequency in the source population. The selected locus is flanked on each side by 100 evenly spaced, selectively neutral polymorphic loci, in analogy to single nucleotide polymorphisms (SNPs) used in genome scans. After the colonization of the novel environments, the derived populations evolve, with each generation including migration from the source population, followed by reproduction with fertility selection and recombination.

Our simulations started with a default parameterization tailored to empirical data from the Ectodysplasin (*Eda*) locus in threespine stickleback, the genomic region where the observation of twin peaks flanking a divergence valley (peak-valley-peak) stimulated our hypothesis of a novel genomic signature of adaptation from shared genetic variation (Roesti *et al.* 2012a). The default model was then expanded to explore the influence of migration rate, time, the number of founder individuals, the strength of divergent selection and recombination rate on the molecular signatures of adaptation, as captured by the magnitude of population divergence (F_{ST} ; Weir & Cockerham 1984) across the neutral loci. We also modified the default model to include *two* selected loci located at equal distances from the centre of the chromosome, which now harboured 400 total neutral loci. Our models first considered divergence between the *source* and the derived populations. These comparisons represent the standard ecological genome scan, as described earlier, and hence served to validate our general simulation approach. In

all subsequent simulations, we focused on divergence among *derived* populations.

Stickleback populations for empirical investigation

Our empirical analyses used marine and freshwater (hereafter M and FW) populations of threespine stickleback. These populations provide an excellent natural system for studying the genomics of parallel adaptation from shared variation because numerous FW populations have been derived independently and recently (<10 000 years ago) from a common M source population (Bell & Foster 1994). Moreover, FW stickleback display relatively consistent phenotypic shifts from their M ancestors as a response to shared selective conditions among FW habitats – that is, parallel adaptation (e.g. Taylor & McPhail 1986; Walker 1997; Walker & Bell 2000; Schluter *et al.* 2004; Berner *et al.* 2010a). Our study considered M stickleback sampled from two estuarine sites on the east coast of Vancouver Island (British Columbia, Canada), and FW samples from a lake and a stream population within each of four independently colonized drainages on Vancouver Island (Berner *et al.* 2008, 2009; Roesti *et al.* 2012a) (Fig. 1; Methods S2, Supporting information). Each of the ten total samples comprised 27 individuals.



Fig. 1 Origin of the stickleback samples used for the empirical analysis. The map shows Vancouver Island (British Columbia, Canada), with the lake and stream populations from four independently colonized freshwater (FW) drainages shown as light and dark coloured circles. The two sites where the marine (M) source population was sampled are shown as black circles.

Targeted sequencing and haplotype networks at M-FW candidate genes and reference loci

We Sanger-sequenced DNA segments at three stickleback candidate genes thought to be under strong divergent selection between M and FW environments. These genes included (i) *Eda*, the major gene underlying the reduction in lateral plate number typically observed when M stickleback colonize FW (Colosimo *et al.* 2005); (ii) *Atp1a1*, a key player in physiological adaptation to osmotically different environments in many organisms (McCormick 2001); and (iii) *Spg1*, which encodes a presumably pH and salinity sensitive glue-like protein used by stickleback males to build their nests (Kawahara & Nishida 2007) (further details on these genes is given in Methods S3, Supporting information). We additionally sequenced a 'reference locus' approximately one megabase away from each of the above three genes. Details on primers and Sanger sequencing are provided in Table S1 and Methods S4 (Supporting information). SNPs derived from these sequences were used to construct haplotype genealogies for each candidate gene and reference locus (Methods S5, Supporting information). We predicted that if adaptation to the replicate derived FW environments at each candidate gene occurred through the parallel fixation of a derived variant present at low frequency in a common M source, all lake and stream samples should form a cluster of closely related haplotypes distinct from the M haplotypes at these loci. Moreover, if M-FW divergence occurred in the face of gene flow, such genealogical structure should not be seen at the three reference loci.

Broad-scale genetic divergence in the candidate regions

To explore divergence at a broader scale around each candidate gene (i.e. across 3–4 Mb 'candidate regions' centred at the genes), we derived SNPs from consensus sequences at genome-wide RAD loci (Baird *et al.* 2008) generated for all 27 individuals from the M and FW samples. (For details on the wet laboratory and consensus genotyping protocols, see Roesti *et al.* 2012a and Methods S6, Supporting information.) These SNPs were used to quantify genetic divergence (F_{ST} based on haplotype diversity; Nei & Tajima 1981 eq. 7) for all possible pairwise comparisons between the two M samples and the eight FW samples (16 total comparisons). We here used only one SNP per RAD locus. We further ensured robust divergence estimation by including a SNP only if both populations in a comparison contributed at least 27 nucleotides to the common nucleotide pool, and if the minor allele frequency across this pool was ≥ 0.25 , thereby eliminating polymorphisms with low information content (Roesti *et al.* 2012b). Moreover,

we corrected for inflated population divergence in chromosome centres relative to their peripheries due to heterogeneous recombination rate along the stickleback chromosomes (Roesti *et al.* 2012a, 2013) by calculating residual divergence (details in Roesti *et al.* 2012a), although qualitatively similar conclusions emerged without this correction. Following these same conventions, we also calculated F_{ST} for pairwise comparisons between the derived FW populations. We here considered comparisons between samples from ecologically similar FW environments only (i.e. six lake–lake and six stream–stream comparisons, 12 in total). The rationale was to avoid capturing selective signatures of lake–stream divergence, which is known to be strong (Berner *et al.* 2008, 2009, 2010b; Deagle *et al.* 2012; Roesti *et al.* 2012a). However, analyses based on *all* possible FW comparisons produced very similar results.

Finally, we generated *overall* M-FW and FW-FW divergence values by averaging residual F_{ST} at each RAD locus across all pairwise M-FW and all pairwise FW-FW comparisons. On average, 6.9 and 6.4 replicate estimates were available per RAD locus for the overall M-FW and FW-FW comparisons, and we achieved a median and mean marker spacing of 12 and 25 kb across the candidate regions (Methods S6, Supporting information). For visualization, we produced sliding window divergence profiles for each candidate region by using the *r* (R Development Core Team 2013) implementation of LOESS (robust locally weighted scatterplot smoothing; Cleveland 1979) (*r* was also used for all other operations unless stated otherwise). We chose a polynomial degree of two and adjusted the smoothing span to achieve equal smoothing resolution across all chromosomes. All genomic positions in this study refer to the improved assembly of the initial (Jones *et al.* 2012b) stickleback reference genome (Roesti *et al.* 2013; <http://data.dryad.org/resource/doi:10.5061/dryad.846nj.2>).

Delta divergence and genealogical sorting in the candidate regions

Parallel divergence between source and derived environments based on shared variation drives opposed patterns of genomic divergence in source-derived versus derived-derived population comparisons (see Results). Calculating the *difference* between overall M-FW and FW-FW divergence, hereafter called ‘delta divergence’, should thus maximize the ability to detect genomic regions underlying parallel divergence (for a proof of principle using simulated data see Fig. S1, Supporting information). We therefore complemented our standard F_{ST} -based divergence analyses described above by generating delta divergence profiles for each

of the three candidate regions (Methods S6, Supporting information).

As an alternative to quantifying genetic divergence between M and FW stickleback based on F_{ST} , we additionally assessed the extent of reciprocal M-FW monophyly captured by phylogenetic trees within the candidate regions. Specifically, we moved a sliding window across the SNPs, calculated a distance matrix for each window, translated each distance matrix to a neighbour joining tree and finally extracted the genealogical sorting index (*gsi*; Cummings *et al.* 2008) from each tree (details in Methods S6, Supporting information). This index ranges from 0 to 1 and quantifies the extent of exclusive ancestry of individuals from defined groups (here M and FW stickleback). The *gsi* data showed a similar physical resolution as the F_{ST} data, and smoothed profiles were generated as described above.

Genome-wide search for signatures of parallel adaptation

To discover additional genomic regions potentially involved in parallel M-FW divergence from shared variation, we performed genome-wide screens of population divergence and genealogical sorting using the RAD-based SNP data and analytical approaches described above. The genome-wide M-FW and FW-FW divergence analyses based on F_{ST} used 16 687 and 16 269 data points (each representing the average of multiple pairwise population comparisons), while the *gsi*-based analysis used 14 890 data points integrating 29 787 phylogenetic trees across the 21 chromosomes. Both types of analyses achieved an approximate genome-wide median and mean marker spacing of 14 kb. We considered a genome region a new candidate if smoothed delta divergence was >0.2 and smoothed *gsi* exceeded 0.6. For each region meeting these criteria, we retrieved all genes located within a window of 400–600 kb centred at the delta divergence peak (generally coinciding exactly with the *gsi* peak) from the Ensembl Genome Browser and assessed whether these genes were known from other (mostly fish) species to be important to saltwater versus freshwater adaptation.

Heterogeneity among chromosomes in M-FW divergence

The presence versus absence of barriers to gene flow (i.e. genes under divergent selection) on specific chromosomes could lead to heterogeneity among chromosomes in the magnitude of population divergence. We considered this possibility by testing for a difference in overall divergence between autosomes under strong

versus weak M-FW selection. The difference between these two chromosome types was defined as those displaying ≥ 2 versus no candidate regions for parallel adaptation from shared variation, as defined in the previous paragraph. For each chromosome, we calculated median raw (not residual, as above) F_{ST} and gsi for each M-FW comparison and averaged these replicate values. We then tested whether the magnitude of divergence differed between the two chromosome categories by permuting F_{ST} and gsi over the chromosomes 9999 times and using the absolute difference between the chromosome categories as test statistic.

Results

Models of parallel adaptation from shared genetic variation

Our models of multiple derived populations diverging from a shared source population into selectively similar environments produced a single peak of high divergence around the selected locus when comparing the source to the derived populations (Fig. 2A). This contrast is the type typically considered in divergence mapping studies. However, our main interest was in *derived-derived* population comparisons, where we found that the parallel fixation of a shared variant leads

to a valley of reduced divergence (hereafter 'divergence valley') around the locus under selection (Fig. 2B). The divergence valley was initially flanked by regions of slightly elevated divergence that then declined towards the chromosome peripheries. In the absence of migration (hence no gene flow), this decline became less striking over time as overall baseline divergence increased owing to drift (Fig. 2B). By contrast, source-derived migration caused the divergence valley to be flanked on either side by striking peaks of high divergence ('Migration' in Fig. 2C). Although these 'divergence twin peaks' emerged consistently across our simulations when comparing derived populations, their height and width were influenced by several factors. First, the peaks grew higher and sharper with increasing time ('Time' in Fig. 2C) and with a decreasing number of founder individuals ('Founders' in Fig. 2C). Second, the physical extent of the divergence twin peaks and of the divergence valley was greater – and could be quite extensive (kilobases to megabases) – when divergent selection was strong and/or recombination was low ('Selection' and 'Recombination' in Fig. 2C).

In our simulations with two loci under divergent selection, separate peak-valley-peak signatures emerged when the selected loci were far apart ('Distant' in Fig. 2D). When the loci were closer together, the entire chromosome segment between them reached high

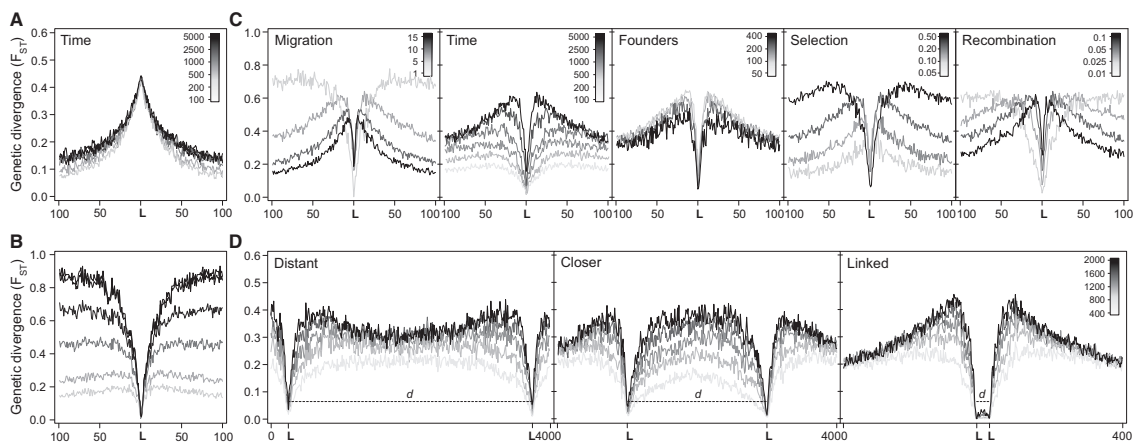


Fig. 2 Molecular signatures of parallel adaptation from shared genetic variation. Shown is the magnitude of population divergence (F_{ST}) at neutral markers along a chromosome segment holding a locus (or loci; L) under divergent selection between a source and multiple derived populations. (A) Traditional genome scan comparing the source to the derived populations, exhibiting a divergence peak at L. (B) By contrast, comparing multiple derived populations adapting in parallel produces a divergence valley around L. In the absence of migration, the rest of the chromosome diverges over time [timescale as in (A)]. (C) Allowing for migration from the source to the derived populations generates a characteristic peak-valley-peak signature of selection ('Migration'). The other panels in (C) show how this signature is influenced by variation in divergence time ('Time', in generations), the number of initial colonizers ('Founders'), the strength of divergent selection on L ('Selection') and the recombination rate across the chromosome segment ('Recombination'). (D) Simulations with two loci occurring at different distances d on the chromosome (the scale indicates divergence time).

divergence ('Closer' in Fig. 2D). Finally, when the selected loci occurred in very close proximity to each other, the two divergence valleys collapsed to a single large valley flanked by particularly pronounced divergence twin peaks ('Linked' in Fig. 2D).

Signatures of parallel adaptation from shared variation at stickleback candidate genes

Haplotype genealogies generated from targeted sequence data at three candidate genes for M-FW adaptation (*Eda*, *Atp1a1*, *Spg1*) consistently revealed the pattern that our simulations suggested should characterize parallel adaptation from shared genetic variation. That is, lake and stream FW samples shared closely related haplotypes that were clearly distinct from the haplotypes predominant in M stickleback ('Candidate gene'

in Fig. 3A). In marked contrast, the reference loci approximately one megabase away from the candidate genes showed little or no habitat-related haplotype structure ('Reference locus' in Fig. 3A). This result indicates the parallel fixation of shared alleles at the candidate genes in FW, despite high M-FW gene flow in other parts of the genome.

We next used SNPs generated through RAD sequencing to assess broad-scale divergence (F_{ST} and genealogical sorting index, gsi) around the three candidate genes for the overall M-FW (source vs. derived) and FW-FW (derived vs. derived) comparison. As expected from the above small-scale targeted Sanger sequencing, M-FW divergence was exceptionally strong close to the three candidate genes (black lines in Fig. 3B), and gsi indicated striking phylogenetic separation between M and FW stickleback in these regions

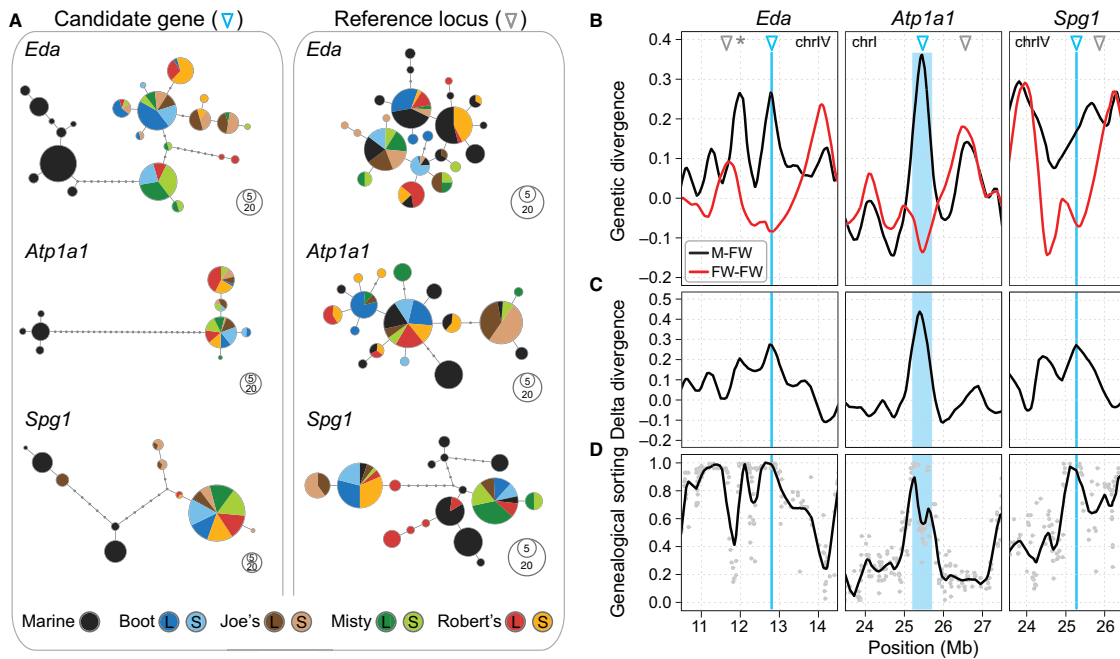


Fig. 3 Genetic structure and divergence at three candidate genes for M-FW divergence in stickleback. (A) Haplotype networks based on targeted sequencing at each candidate gene (left column), and at a corresponding reference locus (right column). The pies represent unique haplotypes, pie sizes reflect their frequency and the internodes are mutational steps. At the candidate genes, lake (L) and stream (S) populations from four independently colonized watersheds form a haplotype cluster separated from the marine (M) haplotype cluster, a genealogical structure not seen (or that is much weaker) at the reference loci. (B) Genetic divergence profiles reveal that the candidate genes (position given by the blue triangle and vertical line; the reference loci are indicated by grey triangles) generally coincide with peaks in M-FW divergence (black) and valleys in FW-FW divergence (red), as expected for parallel adaptation from shared variation. (C) Consequently, delta divergence, calculated as the difference between the two profiles in (B) (i.e. M-FW divergence minus FW-FW divergence; see also Fig. S1) peaks right at the candidate genes. (D) In the same regions, genealogical sorting (gsi) profiles reveal that M and FW stickleback exhibit completely separate ancestry at a broad physical scale (the gsi values underlying the smoothed profile are drawn as grey dots). Note that the neighbourhood of *Eda* contains further regions displaying complete genealogical separation, for instance, at the *Abcb7* gene indicated by a grey asterisk in (B). The location of *Atp1a1* is given as a wide blue vertical line because it is associated with an inversion (Jones *et al.* 2012b).

(Fig. 3D). Importantly, however, the comparison between the *derived* FW populations revealed a valley of low divergence around each candidate gene, as predicted by our simulations (red lines in Fig. 3B). Moreover, these divergence valleys in the FW-FW comparisons were often flanked by striking divergence peaks – some of which were absent in the M-FW comparison, a pattern specifically predicted by our simulations with gene flow. We also found that these signatures of adaptation from shared variation were particularly obvious when M-FW and FW-FW divergence was combined into delta divergence profiles, yielding peaks exactly at the candidate genes (Fig. 3C).

Genome-wide signatures of parallel adaptation from shared variation

We used (delta) divergence and gsi profiles to search the stickleback genome for additional regions likely involved in parallel adaptation from shared variation. This screen discovered 15 such regions on eight chromosomes. Details on these regions, including strong candidate genes for M-FW adaptation (some of which have been suggested previously for stickleback; Hohenlohe *et al.* 2010; Jones *et al.* 2012a,b), are provided in Table S2 (Supporting information), and (delta) divergence and gsi profiles for seven representative new candidate regions are presented in Fig. 4. Full

genome-wide divergence and genealogical sorting profiles are provided in Fig. S2 (Supporting information).

Chromosome-level relationship between candidate regions and divergence

The six autosomes displaying multiple genomic signatures of parallel adaptation from shared variation (i.e. the chromosomes 1, 4, 7, 11, 12, 20) also exhibited exaggerated overall divergence (45% higher F_{ST} on average, $P = 0.0023$; 35% higher gsi , $P = 0.0186$) between M and FW populations compared with the 12 chromosomes lacking such signatures (Fig. 5).

Discussion

We combined simulations and empirical data to shed light on the genomic patterns that arise when multiple populations diverge into selectively similar environments by using shared genetic variation from the ancestral source population. Our main finding is that the immediate neighbourhood of the selected genetic locus underlying parallel adaptation will remain undifferentiated among the derived populations, whereas the broader neighbourhood around the locus will be driven to high divergence. In combination, this produces a characteristic peak-valley-peak signature of genomic divergence among derived populations.

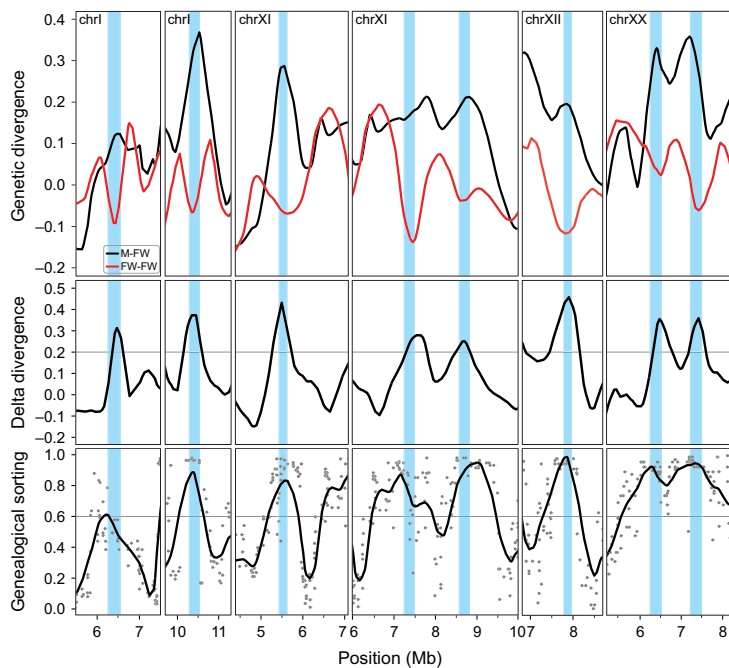


Fig. 4 Exemplary novel regions in the stickleback genome potentially harbouring genes involved in parallel FW adaptation from shared genetic variation. These regions were identified as candidate adaptation hotspots because they displayed high delta divergence (>0.2; threshold shown as grey horizontal line) as a consequence of opposed divergence profiles in M-FW versus FW-FW comparisons, and strong M-FW genealogical sorting (>0.6; grey horizontal line). The plotting conventions follow the ones in Fig. 3B–D. Strong candidate genes located within the blue regions are listed in Table S2.

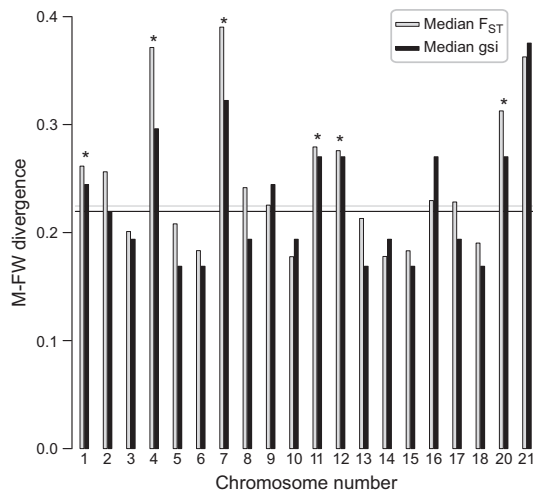


Fig. 5 Heterogeneity among chromosomes in overall divergence (F_{ST} and g_{SI}) between M and FW stickleback. The values represent averages across the 16 replicate population comparisons. Genome-wide median F_{ST} and g_{SI} are given as horizontal grey and black lines. Chromosomes exhibiting two or more signatures of parallel adaptation from shared variation are indicated by an asterisk. Note that the strongly divergent chromosomes (4, 7) are among the three largest ones in the stickleback genome and exhibit particularly low average recombination rates (Roesti *et al.* 2013). The sex chromosome (19) was excluded from this analysis.

Distinct mechanisms drive the peak-valley-peak divergence signature

The peak-valley-peak signature of divergence among derived populations arises from an interaction between two distinct evolutionary mechanisms operating at different timescales. The first mechanism is hitchhiking (Maynard Smith & Haigh 1974; Kaplan *et al.* 1989) of different genomic regions surrounding the genetic variant that is selected in parallel within the derived populations (Fig. 6A). This process halts as soon as the adaptive variant reaches fixation (or some other migration-selection balance) within the derived populations. During this phase, the opportunity for the derived variant to become associated with new genetic backgrounds through recombination is limited. Therefore, in the *close* neighbourhood of the selected locus, the derived populations become fixed for the same haplotype tract linked to the shared adaptive variant. Comparing populations adapted in parallel will thus reveal a low-divergence valley surrounding the locus under common selection (Fig. 6B). In the *broader* neighbourhood of the selected locus, however, recombination during the hitchhiking phase will occur sufficiently frequently to associate the adaptive variant with genetic backgrounds *specific* to

each derived population. These population-specific haplotypes increase in frequency along with the adaptive variant, causing elevated divergence among derived populations on either side of the divergence valley. Even further away from the selected locus, divergence declines again because recombination increasingly associates the derived variant with random haplotypes from the source population. This first hitchhiking phase thus establishes a divergence valley surrounded by a broad region of elevated divergence among the derived populations (Fig. 6B; also see Fig. 2B).

The second mechanism shaping the peak-valley-peak signature is a *long-term* barrier to gene flow caused by the locus under selection (Barton & Bengtsson 1986; Gavrilets & Cruzan 1998; Bierne 2010; Feder & Nosil 2010). Specifically, the selected locus blocks introgression from the source to the derived populations in its chromosomal neighbourhood, while recombination makes introgression increasingly easy with increasing distance from the locus (Fig. 6C). In other words, effective population size is reduced around the selected locus, promoting localized divergence by genetic drift. The divergence valley persists despite elevated drift, however, because the derived populations continue to share the same haplotype tract around the adaptive variant (Fig. 6C). This second mechanism – the barrier to gene flow – thus reinforces and sharpens the broad peak-valley-peak divergence signature that is initiated by hitchhiking.

Determinants of the peak-valley-peak signature

The physical extent of the peak-valley-peak divergence signature is influenced by several factors (Fig. 2C). First, decreasing gene flow between the source and the derived populations causes the peak-valley-peak to become more extensive – because the overall opportunity for introgression decreases. In the extreme case of no gene flow at all, relatively elevated divergence around the divergence valley produced initially by hitchhiking is rapidly eroded because drift causes divergence among the derived populations across the entire chromosome (Fig. 2B). Second, in the presence of gene flow, the peak-valley-peak signature becomes narrower over time as genetic homogenization through introgression moves closer to the selected locus. (Note that a narrower divergence valley is also expected when the derived variant recombines extensively while standing in the source population prior to parallel adaptation; see Discussion S1, Supporting information.) Third, the divergence twin peaks become higher with a decreasing number of individuals founding the derived populations, which increases stochasticity in the haplotypes linked to the derived variant, hence promoting drift. Finally, the peak-valley-peak signature becomes more

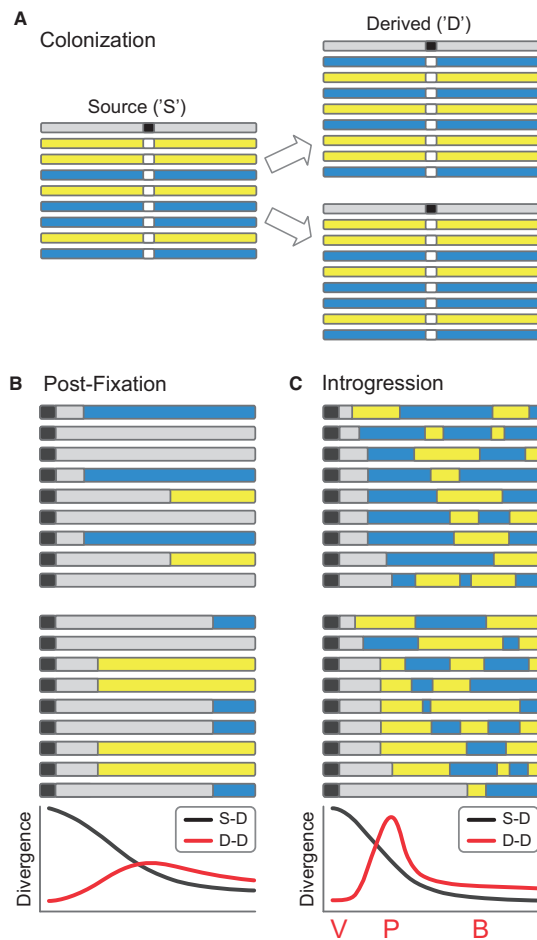


Fig. 6 Mechanisms generating the peak-valley-peak signature of parallel adaptation from shared genetic variation. (A) Multiple novel, selectively similar environments are colonized by a source population occupying a selectively different environment. Individuals are represented by a single haploid chromosome, with different colours indicating different genetic backgrounds. The centre of the chromosome holds a locus under divergent selection, with the white variant favoured in the source population, and the black variant favoured in the derived populations but standing at low frequency in the source population as well. (B) Immediately after the parallel fixation of the selected variant, the derived populations share identical haplotype tracts (grey) near the selected locus, whereas population-specific haplotypes (blue, yellow) have hitchhiked further away from the locus [(B) and (C) show the locus and one side of the chromosome only]. As a result, comparisons between derived populations (red line in the bottom panel) show minimal divergence around the selected locus, flanked by a region of elevated divergence. By contrast, comparisons between the source and derived populations reveal the classical signature of a selective sweep (black line). (C) Continuous migration from the source population causes introgressive hybridization in the derived populations. Introgression is impeded in the neighbourhood of the locus, however, where divergent selection produces a barrier to gene flow that locally promotes population divergence by drift. Consequently, comparisons between derived populations reveal a characteristic genomic signature including a divergence valley ('V' in the bottom panel) caused by haplotype sharing flanked by a divergence peak ('P') reflecting elevated drift. Further away from the locus, population divergence decays to the genome-wide migration-drift baseline level ('B').

extensive with an increasing strength of divergent selection between the source and the derived populations, and with decreasing recombination rate. The reason is that both factors render the barrier to gene flow associated with the selected locus more effective (Barton & Bengtsson 1986; Feder & Nosil 2010).

In our simulations with a single selected locus, the physical extent of the peak-valley-peak divergence signature can be quite substantial – many kilobases to a few megabases. Our two-locus models, however, indicate that even more extensive signatures can emerge when multiple loci are simultaneously under selection. Interestingly, the divergence patterns driven in this latter situation vary qualitatively as a function of the recombination distance between the two loci under selection. When these loci are relatively close to each other, a large region of *high* divergence can arise between them ('Closer' in Fig. 2D), although this region does not hold

either of the selected loci. This pattern arises because the barriers to gene flow associated with the two loci overlap in this region, making introgression particularly difficult. When the selected loci occur in even closer proximity to each other, however, they together bring to fixation a larger genomic segment shared among the derived populations, resulting in a remarkably wide region of *low* divergence ('Linked' in Fig. 2D). Also, the divergence twin peaks flanking this divergence valley are higher than the peaks driven by each locus alone ('Distant' in Fig. 2D), because the two tightly linked loci together drive a single, stronger barrier to gene flow.

Empirical insights from stickleback

Our empirical system provides an appropriate natural analogue for the conditions specified in our simulations. First, no appreciable genetic divergence was present between our two M samples taken 100 km apart (Fig. 1) (median and mean F_{ST} for all pairwise population comparisons are provided in Table S3, Supporting information). This result is consistent with previous reports of very weak genetic structure within M stickleback

(Hohenlohe *et al.* 2010; Jones *et al.* 2012a; Catchen *et al.* 2013), and it generally supports the established idea that present-day M stickleback provide an appropriate surrogate for the ancestors of derived FW populations. Second, haplotype genealogies confirmed that our FW stickleback populations adapted in parallel at three candidate genes involved in M-FW adaptive divergence, specifically by recycling shared variants from a common M source population (see also Colosimo *et al.* 2005; Jones *et al.* 2012b; Deagle *et al.* 2013). Given these results, we scrutinized patterns of genetic divergence around the three candidate genes to empirically test for the signatures of parallel adaptation from shared variation that were suggested by the simulations.

All three candidate regions exhibited the expected genomic signature of parallel adaptation from shared variation: in comparisons between the derived (FW) populations, the selected loci showed low-divergence valleys that were flanked by high-divergence twin peaks. At the same time, classical source-derived (M-FW) comparisons revealed the expected strong divergence at the candidate genes. Combining these opposed FW-FW and M-FW profiles into 'delta divergence' proved a particularly effective way to reveal parallel adaptation from shared variation. One reason is that these profiles reduce heterogeneity in genomic divergence unrelated to a focal ecological factor (here M vs. FW), such as selective sweeps driven by genetic variants favoured in *all* types of habitats (Bierne 2010). An excellent complementary method was to use genealogical sorting in phylogenetic trees (gsi; Cummings *et al.* 2008) to confirm shared ancestry among the FW populations but exclusive ancestry between M and FW populations. Generally, our ability to detect signatures of parallel adaptation from shared variation was greatly enhanced by high replication at the population level. That is, F_{ST} profiles from *single* pairwise population comparisons (M-FW, FW-FW) exhibited substantial noise (details not presented), which would have made interpretations difficult in the absence of multiple such pairs.

The physical scales of the signatures of parallel adaptation from shared variation were extensive – and similar to those suggested by the simulations. For instance, almost full genealogical sorting occurred over several hundred kilobases around each candidate gene. Moreover, the divergence valley around *Eda* was remarkably wide and displayed *two* divergence minima separated by a small rebound in divergence (at ca. 12.5 Mb in Fig. 3B), as well as massive divergence peaks on either side. This pattern strikingly resembles our simulations with two closely linked loci under selection ('Linked' in Fig. 2D). We therefore propose that the broad neighbourhood of *Eda* is influenced by selection on two genes (or gene clusters) that together produce a particularly effective barrier to gene flow from the M source

population. Consistent with this idea, the second divergence minimum near *Eda* coincides with the ATP-binding cassette *Abcb7* (at 12.0 Mb in Fig. 3B), a gene recently suggested to be under divergent selection between M and FW stickleback (Jones *et al.* 2012b). Similarly, a second M-FW adaptation gene near *Spg1* likely influences divergence profiles in that region (Fig. 3B).

A screen of the whole stickleback genome for the joint occurrence of peak-valley-peak signatures of divergence and strong M-FW genealogical sorting identified additional regions on multiple chromosomes likely involved in parallel adaptation from shared variation. As was the case with our initial three candidate genes, these new regions were often flanked by striking divergence twin peaks in the FW-FW comparison, but not in the M-FW comparison (Fig. 4), as predicted by our simulations with gene flow. (Gene flow is known to occur between M and FW populations; Hagen 1967; Jones *et al.* 2006.) Our study thus provides further molecular evidence for divergence in the face of gene flow between contemporary M and FW populations (Catchen *et al.* 2013; Deagle *et al.* 2013). Furthermore, our genome-wide analysis makes a strong case for the notion that adaptation involves numerous loci (e.g. Hohenlohe *et al.* 2010; Lawniczak *et al.* 2010; Fournier-Level *et al.* 2011; Jones *et al.* 2012b; Roesti *et al.* 2012a; Renaut *et al.* 2013), although our methods certainly underestimate the number of loci involved in adaptive divergence between M and FW stickleback (Discussion S2, Supporting information). Finally, our empirical analysis indicated that loci under divergent selection may hinder introgression and drive heterogeneous genomic divergence at the scale of entire chromosomes (Fig. 5).

Implications for ecological genomics

Our results add complexity to the interpretation of regions of low and high divergence discovered in genome scans for signatures of selection. On the one hand, we demonstrate that the common interpretation of regions exhibiting exceptionally low population divergence – that is, localized introgression and balancing selection (Nielsen 2005; Storz 2005) – is potentially problematic; the same pattern can arise when populations use shared genetic variation for parallel adaptation. On the other hand, we also demonstrate that peaks of high population divergence do not necessarily indicate divergent selection. They might instead reflect selectively neutral regions under the influence of neighbouring loci involved in parallel adaptation from shared variation to *similar* environments (for related caveats see Excoffier & Ray 2008; Bierne 2010; Bierne *et al.* 2011). Inference in ecological genomics thus benefits strongly from the integration of multiple complementary analytical

approaches (e.g. source-derived vs. derived-derived comparisons, delta divergence, genealogical sorting; see also Grossman *et al.* 2010), requiring extensive population-level replication within a clear-cut ecological context. On the bright side, genome scans specifically looking for the signature described in our study might help discover adaptation genes in empirical systems where ecological divergence is likely to have occurred repeatedly by recycling genetic variation (e.g. Terai *et al.* 2006; Renaut *et al.* 2011; Tennessen & Akey 2011; Domingues *et al.* 2012; Nadeau *et al.* 2012; Gross & Wilkens 2013).

Acknowledgements

R. Taylor, J.-S. Moore and K. Oke provided marine stickleback samples. A.-C. Grandchamp, J.-S. Moore and K. Hudson aided freshwater stickleback sampling. B. Aeschbach and N. Boileau facilitated wet laboratory work. I. Nissen and C. Beisel performed Illumina sequencing at the Quantitative Genomics Facility, D-BSSE, ETH Zürich. M. Matschiner stimulated the *gsi* analysis and constructed the haplotype networks. N. Bierne and three anonymous reviewers provided valuable comments on the manuscript. M. Hansen, T. Vines and J. Gow speedily processed the manuscript. Financial support came from the National Institute for Mathematical and Biological Synthesis (NSF Award EF-0830858) and the National Institutes of Health (Grant GM56693) (SG); the Natural Sciences and Engineering Research Council of Canada (APH); the European Research Council (Starting Grant INTERGENADAPT, WS); the Swiss National Science Foundation (Sinergia Grant CRSII3_136293, WS; Ambizione PZ00P3_126391/1, DB); the University of Basel (WS, DB), and the Freiwillige Akademische Gesellschaft Basel (DB). We are grateful to all these people and institutions.

References

- Baird NA, Etter PD, Atwood TS, *et al.* (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS One*, **3**, e3376.
- Barrett RDH, Schluter D (2008) Adaptation from standing genetic variation. *Trends in Ecology & Evolution*, **23**, 38–44.
- Barton NH (2000) Genetic hitchhiking. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **355**, 1553–1562.
- Barton N, Bengtsson BO (1986) The barrier to genetic exchange between hybridizing populations. *Heredity*, **57**, 357–376.
- Bell MA, Foster SA (1994) *The Evolutionary Biology of the Threespine Stickleback*. Oxford University, Oxford.
- Berner D, Adams DC, Grandchamp AC, Hendry AP (2008) Natural selection drives patterns of lake-stream divergence in stickleback foraging morphology. *Journal of Evolutionary Biology*, **21**, 1653–1665.
- Berner D, Grandchamp A-C, Hendry AP (2009) Variable progress toward ecological speciation in parapatry: stickleback across eight lake-stream transitions. *Evolution*, **63**, 1740–1753.
- Berner D, Stutz WE, Bolnick DI (2010a) Foraging trait (co)variances in stickleback evolve deterministically and do not predict trajectories of adaptive diversification. *Evolution*, **64**, 2265–2277.
- Berner D, Roesti M, Hendry AP, Salzburger W (2010b) Constraints on speciation suggested by comparing lake-stream stickleback divergence across two continents. *Molecular Ecology*, **19**, 4963–4978.
- Bierne N (2010) The distinctive footprints of local hitchhiking in a varied environment and global hitchhiking in a subdivided population. *Evolution*, **64**, 3254–3272.
- Bierne N, Welch J, Loire E, Bonhomme F, David P (2011) The coupling hypothesis: why genome scans may fail to map local adaptation genes. *Molecular Ecology*, **20**, 2044–2072.
- Catchen J, Bassham S, Wilson T *et al.* (2013) The population structure and recent colonization history of Oregon threespine stickleback determined using RAD-seq. *Molecular Ecology*, **22**, 2864–2883.
- Cleveland WS (1979) Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, **74**, 829–836.
- Colosimo PF, Hosemann KE, Balabhadra S *et al.* (2005) Widespread parallel evolution in sticklebacks by repeated fixation of ectodysplasin alleles. *Science*, **307**, 1928–1933.
- Cummings MP, Neel MC, Shaw KL (2008) A genealogical approach to quantifying lineage divergence. *Evolution*, **62**, 2411–2422.
- Deagle BE, Jones FC, Chan YF *et al.* (2012) Population genomics of parallel phenotypic evolution in stickleback across stream-lake ecological transitions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **279**, 1277–1286.
- Deagle BE, Jones FC, Absher DM, Kingsley DM, Reimchen TE (2013) Phylogeography and adaptation genetics of stickleback from the Haida Gwaii archipelago revealed using genome-wide single nucleotide polymorphism genotyping. *Molecular Ecology*, **22**, 1917–1932.
- Domingues VS, Poh Y-P, Peterson BK *et al.* (2012) Evidence of adaptation from ancestral variation in young populations of beach mice. *Evolution*, **66**, 3209–3223.
- Excoffier L, Ray N (2008) Surfing during population expansions promotes genetic revolutions and structuration. *Trends in Ecology & Evolution*, **23**, 347–351.
- Feder JL, Nosil P (2010) The efficacy of divergence hitchhiking in generating genomic islands during ecological speciation. *Evolution*, **64**, 1729–1747.
- Feder JL, Egan SP, Nosil P (2012) The genomics of speciation-with-gene-flow. *Trends in Genetics*, **28**, 342–350.
- Fournier-Level A, Korte A, Cooper MD *et al.* (2011) A map of local adaptation in *Arabidopsis thaliana*. *Science*, **334**, 86–89.
- Gavrilets S, Cruzan MB (1998) Neutral gene flow across single locus clines. *Evolution*, **52**, 1277–1284.
- Gross JB, Wilkens H (2013) Albinism in phylogenetically and geographically distinct populations of *Astyanax* cavefish arises through the same loss-of-function *Oca2* allele. *Heredity*, **111**, 122–130.
- Grossman SR, Shylakhter I, Karlsson EK *et al.* (2010) A composite of multiple signals distinguishes causal variants in regions of positive selection. *Science*, **327**, 883–886.
- Hagen DW (1967) Isolating mechanisms in threespine sticklebacks (*Gasterosteus*). *Journal of the Fisheries Research Board of Canada*, **24**, 1637–1692.
- Hermisson J, Pennings PS (2005) Soft sweeps: molecular population genetics of adaptation from standing genetic variation. *Genetics*, **169**, 2335–2352.

- Hohenlohe PA, Bassham S, Etter PD, *et al.* (2010) Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLoS Genetics*, **6**, e1000862.
- Jones F, Brown C, Pemberton J, Braithwaite V (2006) Reproductive isolation in a threespine stickleback hybrid zone. *Journal of Evolutionary Biology*, **19**, 1531–1544.
- Jones FC, Chan YF, Schmutz J *et al.* (2012a) A genome-wide SNP genotyping array reveals patterns of global and repeated species-pair divergence in sticklebacks. *Current Biology*, **22**, 83–90.
- Jones FC, Grabherr MG, Chan YF *et al.* (2012b) The genomic basis of adaptive evolution in threespine sticklebacks. *Nature*, **484**, 55–61.
- Kaplan NL, Hudson RR, Langley CH (1989) The 'hitchhiking effect' revisited. *Genetics*, **123**, 887–899.
- Kawahara R, Nishida M (2007) Extensive lineage-specific gene duplication and evolution of the spiggin multi-gene family in stickleback. *BMC Evolutionary Biology*, **7**, 209.
- Lawnczak MKN, Emrich SJ, Holloway AK *et al.* (2010) Widespread divergence between incipient *Anopheles gambiae* species revealed by whole genome sequences. *Science*, **330**, 512–514.
- Maynard Smith J, Haigh J (1974) Hitch-hiking effect of a favorable gene. *Genetics Research*, **23**, 23–35.
- McCormick SD (2001) Endocrine control of osmoregulation in teleost fish. *American Zoologist*, **41**, 781–794.
- Messer PW, Petrov DA (2013) Population genomics of rapid adaptation by soft selective sweeps. *Trends in Ecology & Evolution*, **28**, 659–669.
- Nadeau NJ, Whibley A, Jones RT *et al.* (2012) Genomic islands of divergence in hybridizing *Heliconius* butterflies identified by large-scale targeted sequencing. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **367**, 343–353.
- Nei M, Tajima F (1981) DNA polymorphism detectable by restriction endonucleases. *Genetics*, **97**, 145–163.
- Nielsen R (2005) Molecular signatures of natural selection. *Annual Review of Genetics*, **39**, 197–218.
- Nosil P, Schluter D (2011) The genes underlying the process of speciation. *Trends in Ecology & Evolution*, **26**, 160–167.
- Nosil P, Funk DJ, Ortiz-Barrrientos D (2009) Divergent selection and heterogeneous genomic divergence. *Molecular Ecology*, **18**, 375–402.
- Orr HA (1998) The population genetics of adaptation: the distribution of factors fixed during adaptive evolution. *Evolution*, **52**, 935–949.
- Pritchard JK, Pickrell JK, Coop G (2010) The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. *Current Biology*, **20**, 208–215.
- Renaut S, Nolte AW, Rogers SM, Derome N, Bernatchez L (2011) SNP signatures of selection on standing genetic variation and their association with adaptive phenotypes along gradients of ecological speciation in lake whitefish species pairs (*Coregonus* spp.). *Molecular Ecology*, **20**, 545–559.
- Renaut S, Grassa CJ, Yeaman S *et al.* (2013) Genomic islands of divergence are not affected by geography of speciation in sunflowers. *Nature Communications*, **4**, 1827.
- Roesti M, Hendry AP, Salzburger W, Berner D (2012a) Genome divergence during evolutionary diversification as revealed in replicate lake-stream stickleback population pairs. *Molecular Ecology*, **21**, 2852–2862.
- Roesti M, Salzburger W, Berner D (2012b) Uninformative polymorphisms bias genome scans for signatures of selection. *BMC Evolutionary Biology*, **12**, 94.
- Roesti M, Moser D, Berner D (2013) Recombination in the threespine stickleback genome – patterns and consequences. *Molecular Ecology*, **22**, 3014–3027.
- Schluter D, Clifford EA, Nemethy M, McKinnon JS (2004) Parallel evolution and inheritance of quantitative traits. *American Naturalist*, **163**, 809–822.
- Slatkin M, Wiehe T (1998) Genetic hitch-hiking in a subdivided population. *Genetical Research*, **71**, 155–160.
- Stinchcombe JR, Hoekstra HE (2008) Combining population genomics and quantitative genetics: finding the genes underlying ecologically important traits. *Heredity*, **100**, 158–170.
- Storz JF (2005) Using genome scans of DNA polymorphism to infer adaptive population divergence. *Molecular Ecology*, **14**, 671–688.
- Streisfeld MA, Young WN, Sobel JM (2013) Divergent selection drives genetic differentiation in an R2R3-MYB transcription factor that contributes to incipient speciation in *Mimulus aurantiacus*. *PLoS Genetics*, **9**, e1003385.
- Taylor EB, McPhail JD (1986) Prolonged and burst swimming in anadromous and freshwater threespine stickleback, *Gasterosteus aculeatus*. *Canadian Journal of Zoology*, **64**, 416–420.
- Team RDC (2013) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Tennessen JA, Akey JM (2011) Parallel adaptive divergence among geographically diverse human populations. *PLoS Genetics*, **7**, e1002127.
- Terai Y, Seehausen O, Sasaki T *et al.* (2006) Divergent selection on opsins drives incipient speciation in Lake Victoria cichlids. *PLoS Biology*, **4**, e433.
- Walker JA (1997) Ecological morphology of lacustrine threespine stickleback *Gasterosteus aculeatus* L. (Gasterosteidae) body shape. *Biological Journal of the Linnean Society*, **61**, 3–50.
- Walker JA, Bell MA (2000) Net evolutionary trajectories of body shape evolution within a microgeographic radiation of threespine sticklebacks (*Gasterosteus aculeatus*). *Journal of Zoology*, **252**, 293–302.
- Weir BS, Cockerham CC (1984) Estimating F-statistics for the analysis of population-structure. *Evolution*, **38**, 1358–1370.
- Wu CI (2001) The genic view of the process of speciation. *Journal of Evolutionary Biology*, **14**, 851–865.

M.R. designed the study, carried out the wet laboratory work, analysed and interpreted data and drafted the initial manuscript; S.G. performed the simulations; A.P.H. contributed stickleback samples; W.S. provided infrastructure and financial resources; D.B. designed and directed the study, analysed and interpreted data and wrote the final article, with input from all co-authors.

Data accessibility

Marine and freshwater stickleback RAD sequences: NCBI SRA accession numbers SRP036088 and SRP007695. MATLAB code for the simulations: Supplemental material (Appendix S2). Phased Sanger sequences: Supplemental material (Appendix S3).

Supporting information

Additional supporting information may be found in the online version of this article.

Appendix S1 Containing Methods S1–S6, Discussions S1 & S2, Tables S1–S3, Figures S1 & S2.

Appendix S2 MATLAB code for the simulations.

Appendix S3 Phased Sanger sequences for all candidate genes and their reference loci.

Chapter 4

The genomic signature of parallel adaptation from
shared genetic variation

Supplementary material

Appendices S2 & S3 to Roesti et al. (2014) can be obtained online under:
<http://onlinelibrary.wiley.com/doi/10.1111/mec.12720/supinfo>

SUPPORTING INFORMATION

The genomic signature of parallel adaptation from shared genetic variation

Marius Roesti, Sergey Gavrilets, Andrew P. Hendry, Walter Salzburger, Daniel Berner
(Molecular Ecology, 2014)

Includes:

- Methods S1:** Simulation models of parallel adaptation from shared genetic variation
 - Methods S2:** Stickleback populations for empirical investigation
 - Methods S3:** Stickleback candidate genes for parallel M-FW divergence
 - Methods S4:** Targeted Sanger sequencing at candidate genes and reference loci
 - Methods S5:** Haplotype genealogies for candidate genes and reference loci
 - Methods S6:** Broad-scale analyses around the candidate genes
 - Discussion S1:** How are our theoretical models influenced by more complex haplotype structure around the selected variant?
 - Discussion S2:** Objective and limitations of the genome-wide screen for signatures of parallel adaptation from shared variation
 - Table S1:** Sanger sequencing of the stickleback candidate genes and their reference loci
 - Table S2:** Regions in the stickleback genome identified as candidates for M-FW divergence based on the molecular signature of parallel adaptation from shared variation with gene flow
 - Table S3:** Genome-wide magnitude of divergence in all focal population comparisons
 - Figure S1:** Delta divergence calculated from simulated data
 - Figure S2:** Divergence and genealogical sorting profiles for all autosomes
- Supporting References**

Supporting Methods

Methods S1. Simulation models of parallel adaptation from shared genetic variation

General model of parallel adaptation from shared genetic variation

We developed individual-based models in which multiple populations diverge independently from the same source population into a selectively novel environment. This scenario was inspired by threespine stickleback, a species where numerous populations in freshwater environments have been founded from a common marine source population, but is likely relevant to many other biological systems (e.g., Terai et al. 2006; Renaut et al. 2011; Tennessen & Akey 2011; Domingues et al. 2012; Nadeau et al. 2012; Gross & Wilkens 2013; Streisfeld et al. 2013). For consistency with our empirical study (see below), we model eight derived populations. Individuals are monoecious and represented by a single haploid chromosome. A locus with two alleles under divergent selection between the environments is located in the center of that chromosome. The ancestral allele ‘0’ is favored in the environment of the source population whereas the derived allele ‘1’ is favored in the novel environment. The selected locus is flanked on each side by 100 evenly spaced and selectively neutral loci, in analogy to single nucleotide polymorphisms (SNPs) used in genome scans. Among the n colonizers initially founding each of the derived populations, one individual has a haplotype represented by a uniform sequence of 1’s. The other colonizers and the source population display the ancestral 0 allele at the selected locus and 0 and 1 alleles drawn at random with equal probability at the neutral loci. We thus explicitly assume that the derived allele at the selected locus is initially embedded in a specific genetic background shared among the derived populations (details on this assumption are discussed in the Discussion S1). Because our interest is in the early stages of population divergence and because the freshwater stickleback populations used for our empirical work are young (postglacial), our models ignore novel mutation.

After initial colonization, each derived population grows according to the Beverton-Holt model in non-overlapping generations (Kot 2001). Specifically, the number of offspring produced by each female is taken from a Poisson distribution with parameter $\frac{b}{1 + (b-1)\frac{N}{wK}}$, where b is the

expected number of offspring (set to 10 in all simulations), N is the current population size, K is the environment’s carrying capacity, and w is the female’s fitness. For computational efficiency, we choose $K = 1,000$, emphasizing that additional exploratory simulations with $K = 10,000$ produce similar results supporting identical conclusions. Females with the ancestral 0 or the derived 1 allele have a fitness of $w = 1 - s$ and 1 in the novel environment, where s represents the strength of divergent selection between the environments. Males are assigned to females at random. During

reproduction, the female and male chromosomes recombine. The number of recombination events is drawn independently for each offspring from a Poisson distribution with parameter R . Recombination occurs with uniform probability across the chromosome. In the beginning of each generation, the derived populations each receive N_m migrants from the source population. After t generations, we calculate the magnitude of population divergence (F_{ST} ; Weir & Cockerham 1984) at all neutral loci (we never calculate divergence at the selected locus itself), including all K individuals from all focal populations. (F_{ST} is calculated globally; however, averaging across pairwise population comparisons produced similar results.) The resulting values are averaged across 100 replicate simulations for every parameter combination.

Parameter space and modeling scenarios

The default parameterization of our model is tailored to empirical data from the Ectodysplasin (*Eda*) locus in threespine stickleback, the genomic region where the observation of twin peaks flanking a divergence valley (peak-valley-peak) stimulated our hypothesis of a novel signature of adaptation from shared genetic variation (Roesti et al. 2012a). The default settings include $s = 0.2$ (Barrett et al. 2008), $R = 0.05$ (Roesti et al. 2013), and $t = 5000$ (Bell & Foster 1994). With the default recombination rate of 0.05, the simulated chromosome approximates a 10 - 15 megabase (Mb) segment harboring *Eda* on chromosome IV. Phylogenetic evidence from the *Eda* locus justifies our modeling of the derived allele in a single shared genetic background at the onset of the simulations: present-day freshwater stickleback populations still share nearly identical haplotypes at *Eda*, even across continents (Colosimo et al. 2005; Berner et al. 2010b). We further assume $n = 100$.

Modifications of the default model are used to explore the influence of each parameter on the molecular signatures of adaptation. First, we track population divergence between the *source* population and the derived populations over time ($t = 100, 200, 500, 1000, 2500, 5000$). These comparisons represent the standard ecological genome scan and hence can serve to validate our general simulation approach. In all subsequent simulations, divergence is calculated among the *derived* populations. Here, we first set $N_m = 0$ to study how divergence builds up over time in the absence of gene flow. In reality, however, gene flow will often occur between source and derived populations in the early stages of divergence (Wu et al. 2001; Nosil et al. 2009; Feder et al. 2012). Our main modeling effort is therefore devoted to divergence with gene flow, exploring all possible combinations of N_m (1, 5, 10, 15; default = 5), t (100, 200, 500, 1000, 2500, 5000), n (50, 100, 200, 400), s (0.05, 0.1, 0.2, 0.5), and R (0.01, 0.02, 0.05, 0.1).

Finally, we modify the default model to include two selected loci located at equal distances $d/2$ from the center of the chromosome, which now harbors 400 total neutral loci. The two derived alleles (one per selected locus) beneficial in the derived environment are initially linked (i.e., within a single neutral background), although they rapidly become dissociated by recombination when their frequency in the derived populations is still low. We perform simulations with different values of d (350, 300, 250, 200, 100, 50, 20, 10) and maintain $N_m = 10$ throughout. To achieve a similar overall selection strength as in the single-locus model, we set $s = 0.1$ for each selected locus. Divergence is calculated at $t = 400, 800, 1200, 1600, 2000$. All other parameter values are the same as in the default single-locus model.

Methods S2. Stickleback populations for empirical investigation

Our study uses stickleback samples from two marine ('M') sites and from a lake and stream (freshwater, 'FW') site within each of four independently colonized watersheds (Boot, Joe's Misty, Robert's) on Vancouver Island, British Columbia, Canada (Fig. 1). Sample size was 27 individuals per site. The FW populations are identical to those studied in Roesti et al. 2012a. The M fish were collected with minnow traps from the Cluxewe River estuary (50° 36' 42" N, 127° 11' 02" W) and the Sayward River estuary (location described in Berner et al. 2010a) on the east coast of Vancouver Island. All our estuarine individuals exhibited full plating along their body and a caudal keel, clearly identifying them as M fish (Bell & Foster 1994). In general, marine stickleback are phenotypically highly stable over space and time, exhibit large population sizes, and show little genetic structure over large geographic distances (Bell & Foster 1994; Walker & Bell 2000; Hohenlohe et al. 2010). Present-day marine stickleback are thus considered good surrogates for the ancestor of recently established FW populations (e.g., Walker & Bell 2000; Berner et al. 2010a). Consistent with this view, the two M samples in the present study did not appreciably differ genetically in any of our analyses: first, haplotype data showed no structure between the two M samples (data not shown). Second, median F_{ST} between the two M samples was zero in the genome-wide analysis (mean and median F_{ST} values for all pairwise population comparisons are presented in Table S3). We therefore pooled the two M samples for the haplotype network analysis (Fig. 3).

Methods S3. Stickleback candidate genes for parallel M-FW divergence

To empirically validate the signature of adaptation from shared genetic variation discovered in the simulations, we required loci showing clear signs of parallel divergence in stickleback. We thus focused on three genes suggested to be under strong divergent selection between M and FW

environments. The first candidate gene was *Eda* (Ectodysplasin). M stickleback have a complete plate row along their body, whereas FW populations typically display greatly reduced plating (Bell & Foster 1994). This divergence is thought to primarily reflect differential exposure to predation between the two environments (Reimchen 1992, 1994; Marchinko 2009) and is driven mainly by the repeated fixation of a derived *Eda* allele shared among FW populations (Colosimo et al. 2005, Berner et al. 2010b). Despite selection for the fully plated phenotype (and thus the ancestral M *Eda* allele) in the ocean, individuals heterozygous at *Eda* do still occur at low frequency (Barrett et al. 2008) in the ocean due to recurrent introgression of derived alleles from FW populations (Colosimo et al. 2005; Schluter & Conte 2009). We sequenced a (mainly intronic) 640 bp segment of *Eda* (Table S1).

The second candidate gene was *Atp1a1* (sodium pump subunit alpha-1). This gene is involved in the maintenance of the ion balance and electrolyte homeostasis in different osmoregulatory epithelia (Evans et al. 2005), and has been identified as a physiological key gene in the adaptation to different osmotic environments in many fish species (e.g., stickleback: Hohenlohe et al. 2010; McCairns & Bernatchez 2010; DeFaveri et al. 2011; Shimada et al. 2011; Jones et al. 2012a; killifish: Scott et al. 2004; bull shark: Reilly et al. 2011; brown trout: Larsen et al. 2008; whitefish: Renaut et al. 2011; reviewed in McCormick 2011). We sequenced a (mainly intronic) 380 bp segment of *Atp1a1* (Table S1).

The third candidate gene was *Spg1* (Spiggin). *Spg1* produces a glue-like protein in the kidneys of male stickleback used to stick nesting material together (Wootton 1976; Jakobsson et al. 1999). This glue seems under divergent selection between M and FW environments because of its sensitivity to salinity, pH, and/or temperature (Kawahara & Nishida 2007), and because strong allele frequency shifts between M and FW stickleback have been found at genetic markers in the close neighbourhood of the gene (Hohenlohe et al. 2010; DeFaveri et al. 2011; Shimada et al. 2011). We sequenced a 356 bp segment of *Spg1* (Table S1). This segment was intergenic but directly adjacent to one of the *Spg1* gene copies.

For each of the three candidate genes, we performed Sanger sequencing (see Methods S4), screened these sequences for polymorphisms, and derived haplotype networks (see Methods S5). We then followed the same steps to Sanger sequence an additional ‘reference locus’ (mainly intergenic, length ranging from 326 – 767 bp) approximately one megabase away from each candidate gene. We predicted that if adaptation to the replicate derived FW environments at each candidate gene occurred through the parallel fixation of a derived variant present at low frequency in a common M source, all lake and stream samples should form a cluster of closely related

haplotypes distinct from the M haplotypes at these loci. Moreover, if M-FW divergence occurred in the face of gene flow, such genealogical structure should not be seen at the three reference loci.

Methods S4. Targeted Sanger sequencing at candidate genes and reference loci

PCR amplification primers for the three candidate genes and their associated reference loci (i.e., six total DNA segments) were designed based on the improved assembly (Roesti et al. 2013) of the stickleback reference genome (Jones et al. 2012b), and based on RAD sequences available from previous work (Roesti et al. 2012a). The primer sequences and amplification conditions are provided in Table S1. The resulting sequences were read on an ABI3130xl capillary sequencer (Applied Biosystems). Each sequence was run at least twice for each individual, usually with both the forward and reverse primer. This allowed unambiguously identifying the diploid genotype of each individual at each candidate gene and reference locus. On average, each candidate and reference locus was sequenced in 64 FW individuals (128 haplotypes), averaging eight fish per FW sample, and in 23 M individuals (46 haplotypes), including fish from both M samples.

Methods S5. Haplotype genealogies for candidate genes and reference loci

To construct haplotype genealogies for the candidate genes and reference loci, we first used CodonCode Aligner v.3.5.6 (CodonCode Corporation) to call diploid consensus sequences and to find SNPs. All polymorphisms were then concatenated (treating indels as a single mutational steps) and phased using PHASE 2.1 (Stephens et al. 2001; Stephens & Donnelly 2003), optimizing the procedure by specifying the polymorphisms' physical positions. Finally, we used jModelTest v0.1.1 (Posada 2008) to identify GTR as the best model of sequence evolution for all polymorphisms, used the maximum-likelihood method implemented in PAUP* v4.0 (Swofford 2003) to determine the most probable genealogical relationship among all individuals at each of the six loci, and visualized these haplotype genealogies following Salzburger et al. (2011).

Methods S6. Broad-scale analyses around the candidate genes

To generate broad-scale profiles of divergence and genealogical structure around the three candidate genes, we used consensus sequences from genome-wide RAD (Baird et al. 2008) loci previously generated for all 27 individuals from each of the eight FW samples (details on the wet lab and consensus genotyping protocols are given in Roesti et al. 2012a). We also generated new, comparable RAD data for the M samples based on the same wet lab protocol, with just two modifications: the final library amplification was performed in seven replicate PCRs to reduce amplification variance, and all 54 M individuals were single-end sequenced on a single Illumina

HiSeq lane with 100 cycles. For the M individuals, consensus genotype sequences at the RAD loci were called as in Mateus et al. (2013). After combining the consensus sequences across all M and FW individuals, each RAD locus was screened for SNPs, including a small fraction of micro-indels. All genomic positions in this study refer to the reference genome re-assembly of Roesti et al. (2013).

SNPs in the three ‘candidate regions’, defined as a 3 - 4 Mb segment around each gene, were used to quantify genetic divergence between M and FW stickleback (F_{ST} based on haplotype diversity; equation 7 in Nei & Tajima 1981). Divergence was calculated for all possible pairwise comparisons between the two M samples and the eight FW samples (16 total comparisons). Robust divergence estimation was ensured by including a SNP only if both populations in a comparison contributed at least 27 nucleotides to the common nucleotide pool, and if the frequency of the minor allele across the nucleotide pool was at least 0.25. The latter criterion eliminated polymorphisms with low information content (Roesti et al. 2012b). In addition, we used only one SNP per RAD locus. Following these same conventions, we then calculated F_{ST} for pairwise comparisons among the derived FW populations. We here considered comparisons among samples from ecologically similar FW environments only (i.e., six lake-lake and six stream-stream comparisons, for 12 comparisons in total). The rationale for excluding lake-stream comparisons was to avoid capturing selective signatures of lake-stream divergence, which is known to be strong (Berner et al. 2008, 2009; Deagle et al. 2012; Roesti et al. 2012a). However, analyses based on *all* possible FW comparisons produced very similar results supporting identical conclusions.

The interaction between selection and heterogeneous recombination rate along stickleback chromosomes can inflate population divergence in chromosome centers relative to their peripheries (Roesti et al. 2012a, 2013). Correcting for this effect by calculating *residual* divergence facilitates the search for signatures of selection (details given in Roesti et al. 2012a). This correction was performed here, although qualitatively similar conclusions emerged either way. Finally, to obtain overall M-FW and FW-FW divergence profiles, we averaged divergence estimates at each RAD locus (residual F_{ST} values) across all pairwise M-FW and all pairwise FW-FW comparisons. This procedure yielded, on average, 6.9 and 6.4 replicate values per RAD site for the overall M-FW and FW-FW contrast. For the *Eda* candidate region (4 Mb in size), the final resolution was 178 and 168 data points for the overall M-FW and FW-FW comparison. The corresponding values for *Atp1a1* (4 Mb) were 193 and 187, and for *Spg1* (3 Mb) 106 and 100. Thus, the median and mean marker spacing in the candidate gene and control regions was 12 and 25 kb respectively (treating markers on sister RAD loci as individual data points).

Parallel divergence between source and derived environments based on shared variation

drives a divergence peak close to the selected locus in source-derived comparisons, but a valley in derived-derived comparisons (see Results). Calculating the *difference* between overall M-FW and FW-FW divergence, hereafter called ‘delta divergence’, should thus maximize the ability to detect genomic regions underlying parallel divergence (for a proof of principle using simulated data, see Fig. S1). We therefore complemented our standard divergence analyses described above by creating delta divergence profiles for each candidate region. We first averaged overall M-FW and FW-FW divergence separately across non-overlapping 5 kb windows, and then, for each window, we subtracted the resulting FW-FW value from its M-FW counterpart. Working with windows enhanced the power of this analysis because divergence data from *both* the M-FW and FW-FW comparison were not available from all RAD loci.

As a complementary approach to quantifying genetic divergence between M and FW stickleback, we assessed the extent of reciprocal M-FW monophyly captured by phylogenetic trees within the candidate regions. Specifically, we moved a sliding window across the SNPs and, for each window, calculated a distance matrix based on the ‘F84’ nucleotide substitution model (Felsenstein 1984). We here accepted multiple SNPs on a RAD locus and used a window size of 33 SNPs, which was the smallest number of markers consistently allowing distance matrix calculation across all windows. The genomic position of a window was defined as the RAD locus position of its central SNP. The distance matrices were then translated to midpoint-rooted neighbor joining trees, which in turn allowed calculating the genealogical sorting index (gsi; Cummings et al. 2008). This index ranges from 0 to 1 and quantifies the extent of exclusive ancestry of individuals from defined groups (here M and FW stickleback) in a phylogenetic tree. If multiple gsi values were available for a RAD locus (owing to multiple SNPs at that locus), they were averaged to a single data point. This analysis yielded 167, 178, and 103 gsi values for the *Eda*, *Atp1a1*, and *Spg1* candidate regions, thus resulting in a similar physical resolution as the F_{ST} -based divergence analysis. The gsi analysis was performed using the R (R Development Core Team 2013) packages APE (Paradis et al. 2004) and genealogicalSorting (<http://www.genealogicalsorting.org>).

Supporting Discussions

Discussion S1. How are our theoretical models influenced by more complex haplotype structure around the selected variant?

Our simulations assume that the genetic variant adaptive in the derived populations has a *single* origin and is thus initially embedded in a single genetic background in all derived populations. Indeed, for the *Eda* locus inspiring our theoretical analysis, phylogenetic data have amply demonstrated extensive sharing among multiple FW populations of the *same* haplotype linked to a derived variant (Colosimo et al. 2005; top left haplotype network in Fig. 3A in this study). This strongly suggests a *single* origin of the derived variant (Colosimo et al. 2005). Our phylogenetic data from the two other candidate genes further indicate that this conclusion is not restricted to *Eda* (see Fig. 3A, middle and bottom left haplotype networks). Indeed, whole-genome re-sequencing supports the view that most of the genetic variation used for parallel FW adaptation has a common origin (Jones et al. 2012b). Extensive haplotype sharing among the derived populations is thus an adequate assumption in our models. We also highlight that with the simulation parameters chosen, the ‘chromosome’ in our models actually corresponds to a relatively narrow segment of a (stickleback) chromosome only.

Nevertheless, it should be kept in mind that the opportunity for a derived FW-adaptive variant to segregate in the M source population prior to selection (and hence to recombine into M genetic backgrounds) will influence the signature of parallel adaptation. Specifically, recombination of the derived variant in the source population will reduce the physical extent of haplotype sharing around the derived variant, eventually causing a *more narrow* divergence valley among the derived populations. This effect is analogous to the erosion of genetic divergence around a selected locus observed in soft sweep models focusing on ancestral versus derived populations (e.g., Hermisson & Pennings 2005; Barrett & Schluter 2008; Messer & Petrov 2013). We can thus make the qualitative predication that loci under strong divergent selection should exhibit a wider divergence valley than weakly selected loci. The reason is that in the former case, a derived variant introduced from a derived population back into the source population by hybridization will be eliminated relatively rapidly from the source population, thus reducing the opportunity for recombination. Similarly, loci situated in low-recombination regions of the genome should exhibit wider divergence valleys (see Discussion S2). We emphasize, however, that the emergence of the flanking *divergence twin peaks* is unaffected by the extent of haplotype sharing around the derived variant (see Fig. 6).

While our models assume a single origin of the derived variant, parallel adaptation among populations can also be based on *multiple* genetic variants produced independently by mutation

(Barrett & Schluter 2008; Messer & Petrov 2013). This scenario was not the focus of our investigation because parallel M-FW divergence from repeated *de novo* mutation is certainly not frequent in stickleback (see above). We can nevertheless make the qualitative predictions that, first, the availability of several independent derived variants in the source population prevents the emergence of a divergence valley among derived populations. This is because the selective sweeps will bring distinct haplotypes to fixation among the derived populations. Second, as above, the emergence of high divergence around the selected locus should still be observed because the barrier to gene flow mechanism operates irrespectively of the initial haplotype structure around the derived variant.

Discussion S2. Objective and limitations of the genome-wide screen for signatures of parallel adaptation from shared variation.

The goal of our genome-wide analysis was primarily to illustrate how the signature of parallel adaptation from shared variation can serve as a tool for the genome-wide detection of genes or chromosome regions involved in parallel adaptation – we did not attempt a complete quantitative investigation of the genetic architecture of M-FW divergence in stickleback. A first limitation is that our RAD marker data lack the physical (basepair) resolution to determine whether a selective signature is driven by a single gene, as opposed to multiple genes clustered within a few kb (our median and mean marker spacing is 14 kb, considering both sister tags associated with a restriction site). Nevertheless, our study exhibits an unprecedented *biological* resolution, as we include 8 FW and 2 M population samples, each represented by 27 individuals (Table S3). Overall FW-FW and M-FW divergence estimates at our SNP markers are thus exceptionally robust.

The power of detecting parallel adaptation regions is further complicated by heterogeneous recombination rate. As our models show, the genomic signature of adaptation from shared variation becomes physically more extensive (and hence easier to detect given limited marker resolution) with decreasing recombination rate ('Recombination' in Fig. 2C). Since recombination rate is much higher in the stickleback chromosome peripheries than in chromosome centres (Roesti et al. 2012a, 2013), we certainly overlook small-scale selective signatures in the chromosome peripheries. Although not widely appreciated, this bias potentially also affects other types of genome-wide scans relying on linkage (Roesti et al. 2013).

Furthermore, the selective signature at a locus is weakened when the same FW-beneficial variant is used for adaptation in a *subset* of the replicate FW populations only. This may occur because this allele simply failed to invade some FW watersheds, or because an adaptive phenotypic

change was achieved in some populations through a different genetic pathway. Like previous genomic analyses in the species (Hohenlohe et al. 2010; Jones et al. 2012a,b), our genome-wide screen is thus biased toward discovering signatures of M-FW divergence caused by alleles recycled *with high fidelity* among FW populations.

Finally, if a variant adaptive in FW managed to recombine effectively into diverse M genetic backgrounds prior to selection, we expect a narrow divergence valley only (see Discussion S1). Given relatively coarse marker resolution, this locus might thus escape our screen for the full signature of parallel adaptation from shared variation introduced in this paper (divergence valley *and* twin peaks). For all these reasons, the candidate regions identified in our genome-wide screen certainly represent only a subset of the M-FW adaptation genes in our study populations.

Supporting Tables

Table S1. Sanger sequencing of the stickleback candidate genes and their reference loci

Amplification PCR reaction volume was 12.5 μ l, with 1 μ l of genomic DNA (concentration: 20 ng/ μ l) using RedTaq (Sigma-Aldrich) (default) or AmpliTaq (Applied Biosystems) polymerase. The following cycling conditions were used for PCR amplification: 1 x 94 °C for 3 min; followed by 30 x 94 °C for 30 sec, X °C for 45 sec and 72 °C for 45 sec; followed by 1 x 72 °C for 7 min and finally hold at 4 °C. Annealing temperatures (X) for particular primer pairs were (in °C): A/B=52.0, C/D=52.0, E/F=53.0, E/G=55.0, H/I=53.0, J/K=54.0, L/M=51.5, N/O=51.0, P/Q=53.5. Each PCR product was then purified by following the ExoSAP-IT (Affymetrix) standard protocol. For the sequencing PCR, we used the BigDye Terminator v3.1 Cycle Sequencing Kit (Applied Biosystems) and added 0.5 μ l primer (forward or reverse) and 1.0 μ l BigDye Terminator Reaction Mix to each purified PCR product. The conditions for the subsequent sequencing PCR were: initial denaturation (1 min, 94 °C) followed by 25 cycles of denaturation (10 s, 94 °C), annealing (20 s, 52 °C) and elongation (4 min, 60 °C). Unincorporated BigDye terminators were removed with the BigDye XTerminator™ Purification Kit (Applied Biosystems), by adding 14.5 μ l ddH₂O, 22.5 μ l SAM™ solution and 5.0 μ l XTerminator™ beads to the sequencing products. After shaking for 30 min at 2000 rpm, the mix was centrifuged (2 min, 1500 rpm).

ID	Gene	Region (inferred from FW-FW comparisons using RADseq)	Chromosome	Genomic position (BROAD S1, Feb 2006; Database version: 68.1)	Genomic position (assembly-corrected reference; Roesti et al. 2013)	Direction	Primer sequence	Tm (°C)
A	<i>Eda</i>	candidate	4	12808304 - 12808322	12808304 - 12808322	Forward	5'-GAC TGG AAG GGA AAG AAG G-3'	55.3
B	<i>Eda</i>	candidate	4	12807800 - 12807820	12807800 - 12807820	Forward	5'-CTG CGC ACA GAG CGT AAA CAC-3'	59.5
C	<i>Eda</i>	candidate	4	12807758 - 12807775	12807758 - 12807775	Forward	5'-CAC AAG AGC AGC GAG ACG-3'	57.1
D	<i>Eda</i>	candidate	4	12809097 - 12809122	12809097 - 12809122	Reverse	5'-GGA ATT CTT TTG TTT TTG TCT TAT CG-3'	53.4
E	<i>Eda</i>	candidate	4	12808548 - 12808569	12808548 - 12808569	Reverse	5'-CCT GTG AAG AGC GAA AGC AAA G-3'	57.7
F	<i>Eda</i>	reference	4	11721754 - 11721770	11721754 - 11721770	Forward	5'-CAT CCA GGC CCA CAA TC-3'	54.0
G	<i>Eda</i>	reference	4	11722645 - 11722664	11722645 - 11722664	Reverse	5'-GGC CTC ATT ACA TAC ATT GC-3'	53.4
H	<i>Atp1a1</i>	candidate	1	21725072 - 21725091	25487439 - 25487458	Forward	5'-GTG TTT ACT CAA GGG AGA GG-3'	55.4
I	<i>Atp1a1</i>	candidate	1	21724269 - 21724286	25488244 - 25488261	Reverse	5'-CAG TCC AAC CTG CCC ATC-3'	57.1
J	<i>Atp1a1</i>	reference	1	20601250 - 20601272	26611258 - 26611280	Forward	5'-GAG CTT TTA TAC GTC TCT GAA GG-3'	55.8
K	<i>Atp1a1</i>	reference	1	20601933 - 20601953	26610577 - 26610597	Reverse	5'-CAC CTC AGT AGG ACA GAA AGC-3'	57.4
L	<i>Spg1</i>	candidate	4	21189272 - 21189296	24986448 - 24986427	Forward	5'-GCT GAG TAC AAT GTT TTA TAT AAC C-3'	53.0
M	<i>Spg1</i>	candidate	4	21189764 - 21189788	24985956 - 24985980	Reverse	5'-CTA CGA ATC TAG AAA TTG TAA GAA G-3'	53.0
N	<i>Spg1</i>	reference	4	20191013 - 20191032	25984712 - 25984731	Forward	5'-GCT TTA GAT TTC ATC GGG AG-3'	53.4
O	<i>Spg1</i>	reference	4	20191548 - 20191566	25984178 - 25984196	Reverse	5'-CAT CAG TAT CTG GCT TTG G-3'	52.9
P	<i>Spg1</i>	reference	4	20190803 - 20190821	25984923 - 25984941	Forward	5'-CGA AGG CCG AAG TTT AAG G-3'	55.3
Q	<i>Spg1</i>	reference	4	20191323 - 20191343	25984401 - 25984421	Reverse	5'-CTT CTG AAA CGT CCG CTT ATG-3'	55.8

Table S2. Regions in the stickleback genome identified as candidates for M-FW divergence based on the molecular signature of parallel adaptation from shared variation with gene flow

A genomic region qualified as M-FW candidate if smoothed delta divergence reached at least 0.2 and smoothed gsi was at least 0.6 (see Figure S2). The last column lists strong candidate genes for M-FW divergence contained in these regions, based on evidence from studies in stickleback (references with double asterisk) and/or other (mostly fish) species (references with single asterisk). Some of these candidate regions are visualized in Figure 4.

Chromosome	Novel candidate region in Mb, corrected for misassembly according to Roesti et al. (2013) (uncorrected in parentheses)	Candidate genes
1	6.3 - 6.6 (6.3 - 6.6)	<i>Tyr</i> (Koga et al. 1995*, Hoegg et al. 2004*; Page-McCaw et al. 2004*; Greenwood et al. 2011**)
1	10.2 - 10.6 (10.2 - 10.6)	<i>Sucnr1</i> (Deen & Robben 2011*)
4	7.3 - 7.7 (7.3 - 7.7)	n.a.
7	10.6 - 11.00 (chrUn 7.95 - 8.35)	n.a.
7	19.2 - 19.6 (17.45 - 17.85)	n.a.
7	19.6 - 20.0 (17.85 - 18.25)	<i>Ncc</i> (Inokuchi et al. 2008*; Shimada et al. 2011**)
8	8.6 - 9.0 (8.6 - 9.0)	<i>Adams10</i> (Hohenlohe et al. 2010**)
9	13.1 - 13.5 (8.7 - 9.1)	n.a.
11	5.3 - 5.8 (5.3 - 5.8)	<i>Atp6v0a1</i> (Hohenlohe et al. 2010**, Jones et al. 2012b**); <i>Kcnh4</i> (Hohenlohe et al. 2010**, Jones et al. 2012b**); <i>Stat3</i> (Hohenlohe et al. 2010**); <i>Fzd2</i> (Hohenlohe et al. 2010**)
11	7.3 - 7.8 (7.3 - 7.8)	<i>Slc35b1</i> (Dejima et al. 2009*); <i>Ndufa4</i> (van Rooijen et al. 2009*)
11	8.4 - 8.9 (8.4 - 8.9)	<i>Kcnj2</i> (= <i>Kir2.1</i>) (Malinowska et al. 2003*); <i>Oto2</i> (Wang et al. 2011*)
12	6.8 - 7.2 (12.38 - 12.78)	n.a.
12	7.5 - 8.1 (11.52 - 12.12)	n.a.
20	6.2 - 6.7 (12.38 - 12.88)	n.a.
20	7.2 - 7.6 (11.39 - 11.89)	n.a.

Table S3. Genome-wide magnitude of divergence in all focal population comparisons

Divergence is expressed as median F_{ST} (mean F_{ST} in parentheses) calculated across all SNPs. For details on the stringent SNP filtering conventions applied to maximize the robustness of divergence estimation see main text.

Ecological contrast	Population 1	Population 2	F_{ST}
M-M	Sayward	Cluxewe	0.0000 (0.0230)
M-FW	Sayward	Boot lake	0.2711 (0.3147)
M-FW	Sayward	Robert's lake	0.0697 (0.1817)
M-FW	Sayward	Joe's Lake	0.3900 (0.4175)
M-FW	Sayward	Misty lake	0.1553 (0.2663)
M-FW	Sayward	Boot stream	0.3690 (0.3718)
M-FW	Sayward	Robert's stream	0.1546 (0.2453)
M-FW	Sayward	Joe's Stream	0.3919 (0.4201)
M-FW	Sayward	Misty Stream	0.1521 (0.2657)
M-FW	Cluxewe	Boot lake	0.2541 (0.2939)
M-FW	Cluxewe	Robert's lake	0.0965 (0.1932)
M-FW	Cluxewe	Joe's lake	0.3637 (0.3848)
M-FW	Cluxewe	Misty lake	0.1715 (0.2449)
M-FW	Cluxewe	Boot stream	0.2826 (0.3438)
M-FW	Cluxewe	Robert's stream	0.1715 (0.2449)
M-FW	Cluxewe	Joe's stream	0.3589 (0.3899)
M-FW	Cluxewe	Misty Stream	0.1765 (0.2588)
FW-FW	Boot lake	Robert's lake	0.2658 (0.3011)
FW-FW	Boot lake	Joe's lake	0.5158 (0.5276)
FW-FW	Boot lake	Misty lake	0.3647 (0.3844)
FW-FW	Robert's lake	Joe's lake	0.3511 (0.3630)
FW-FW	Robert's lake	Misty lake	0.1277 (0.2058)
FW-FW	Joe's lake	Misty lake	0.3706 (0.3902)
FW-FW	Boot stream	Robert's stream	0.3896 (0.3932)
FW-FW	Boot stream	Joe's stream	0.5271 (0.5420)
FW-FW	Boot stream	Misty stream	0.4053 (0.4025)
FW-FW	Robert's stream	Joe's stream	0.4065 (0.4231)
FW-FW	Robert's stream	Misty stream	0.2176 (0.2699)
FW-FW	Joe's stream	Misty stream	0.3816 (0.3957)

Supporting Figures

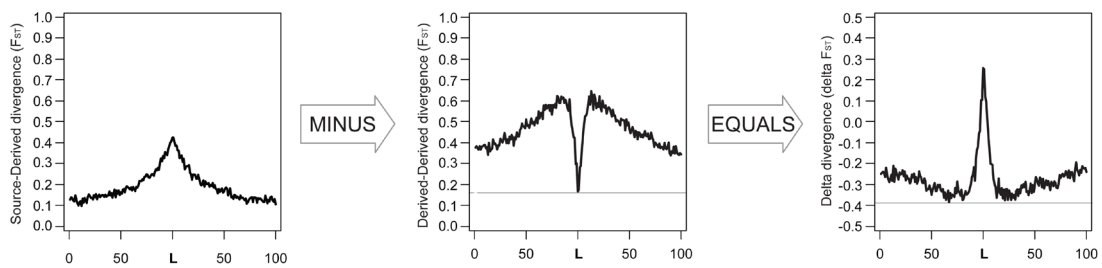


Figure S1. Delta divergence calculated from simulated data

The rationale for using delta divergence to identify genomic regions involved in parallel adaptation from shared variation, illustrated using simulated data generated by the default model (Fig. 2C). Delta divergence is calculated by subtracting the divergence among derived populations (i.e., overall FW-FW divergence in our study) from the divergence between source and derived populations (overall M-FW divergence). The benefit is that the resulting delta divergence peak is higher and sharper than the source-derived peak and the derived-derived valley.

Figure S2. Divergence and genealogical sorting profiles for all autosomes (presented on the 10 pages that follow)

Genetic divergence (based on residual F_{ST} ; see Materials and Methods) between M and FW stickleback populations (top panel, black line) and among FW populations (top panel, red line), resulting delta divergence (middle panel), and M-FW genealogical sorting (bottom panel) plotted for all autosomes. Plotting conventions are as in Fig. 3 (B - D) and Fig. 4.

Figure S2 (chromosomes 1 and 2)

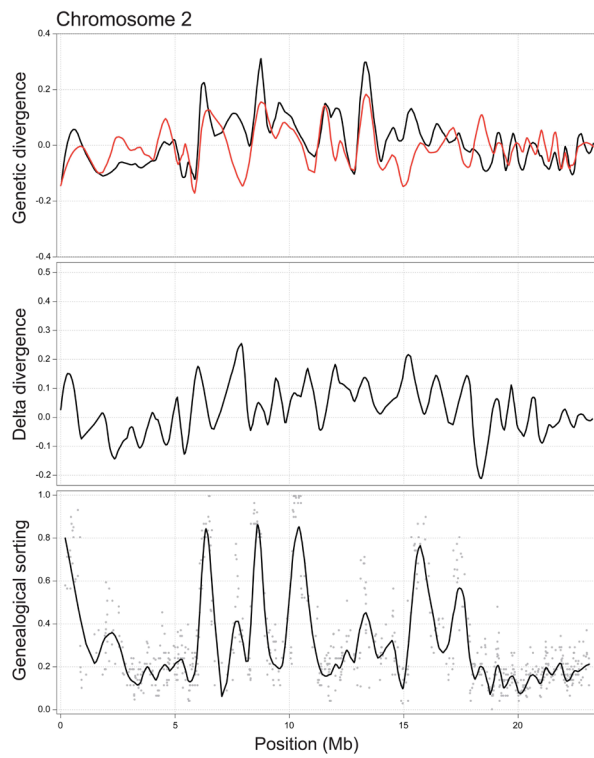
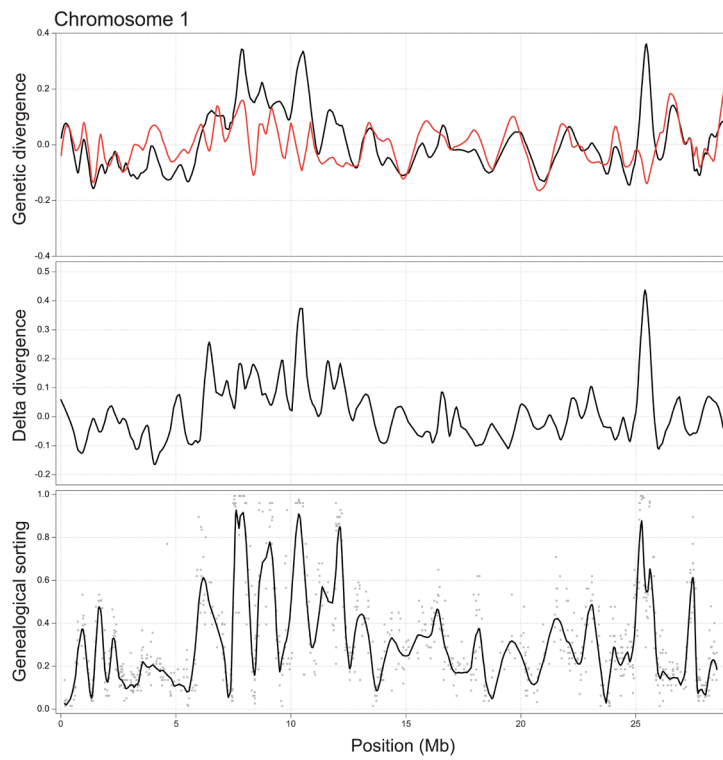


Figure S2 (chromosomes 3 and 4)

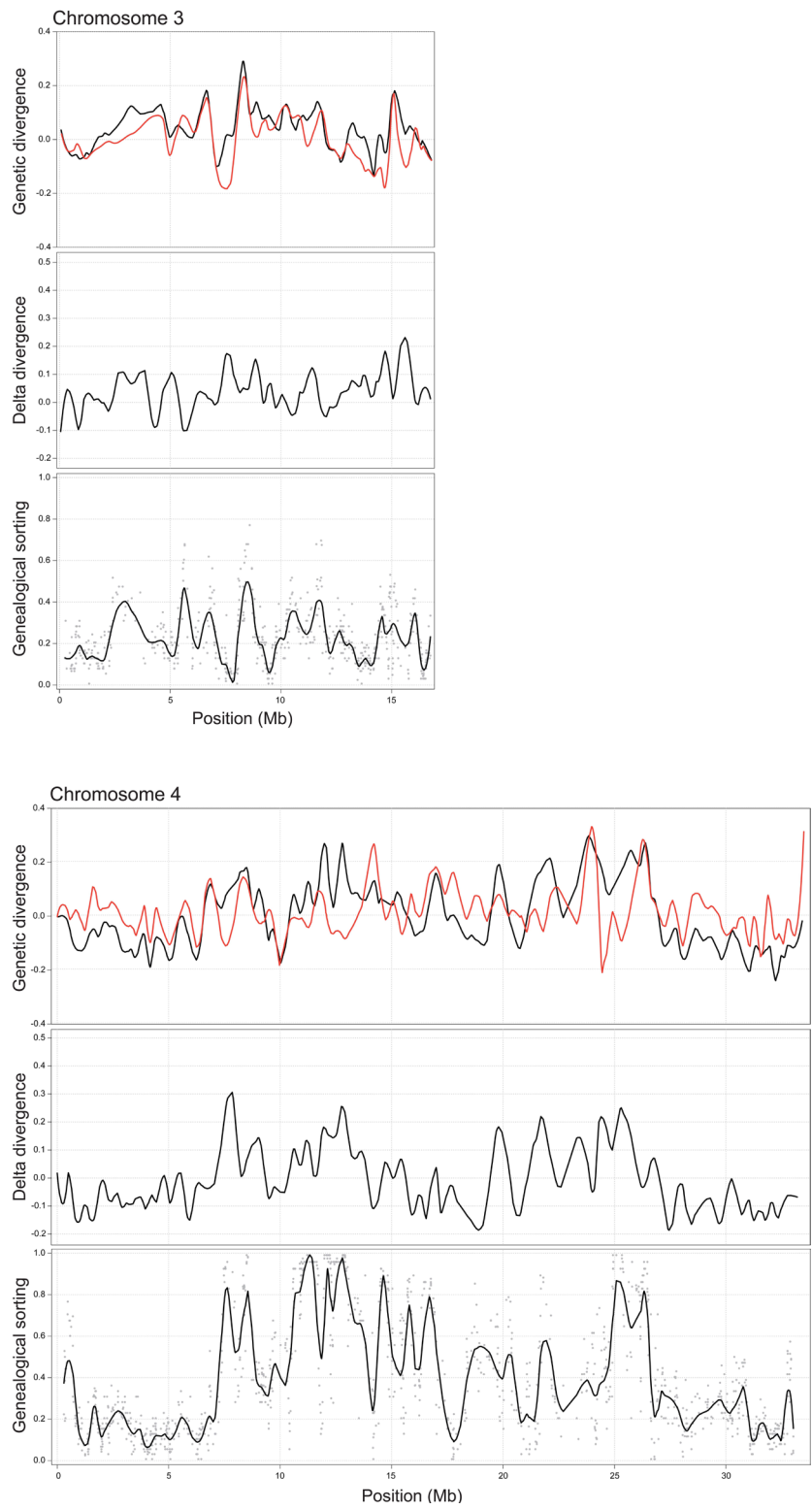


Figure S2 (chromosomes 5 and 6)

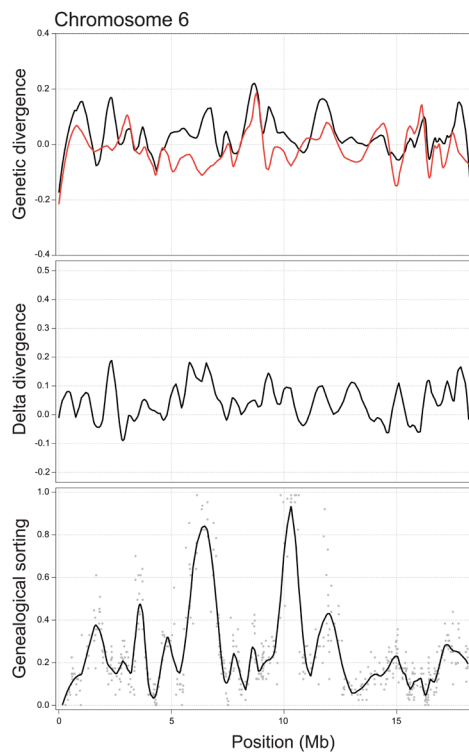
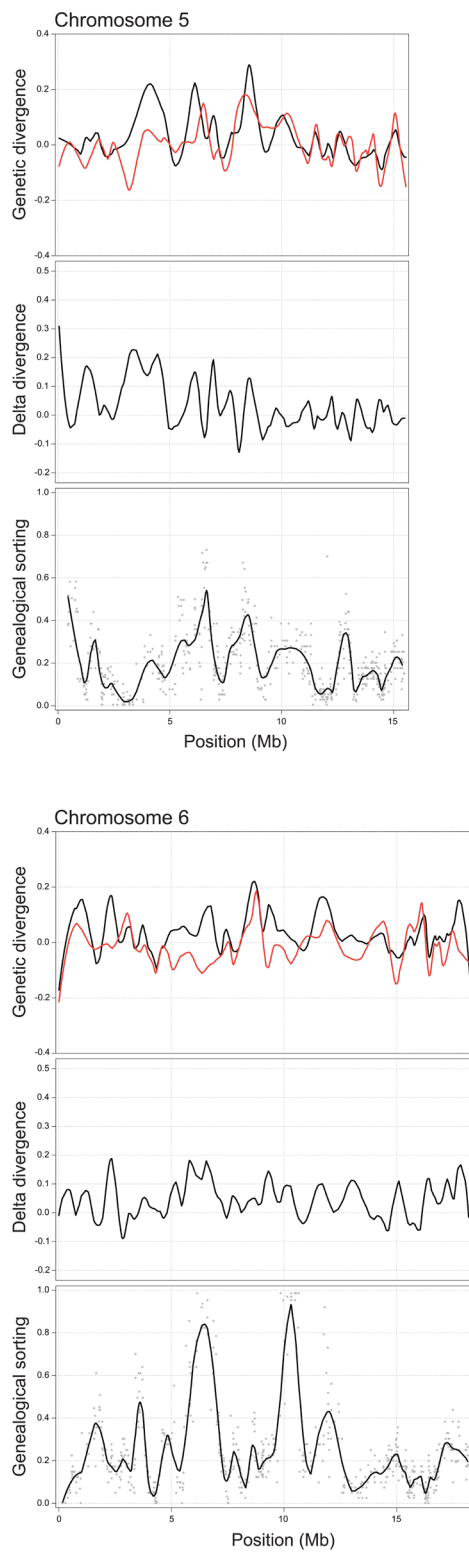


Figure S2 (chromosomes 7 and 8)

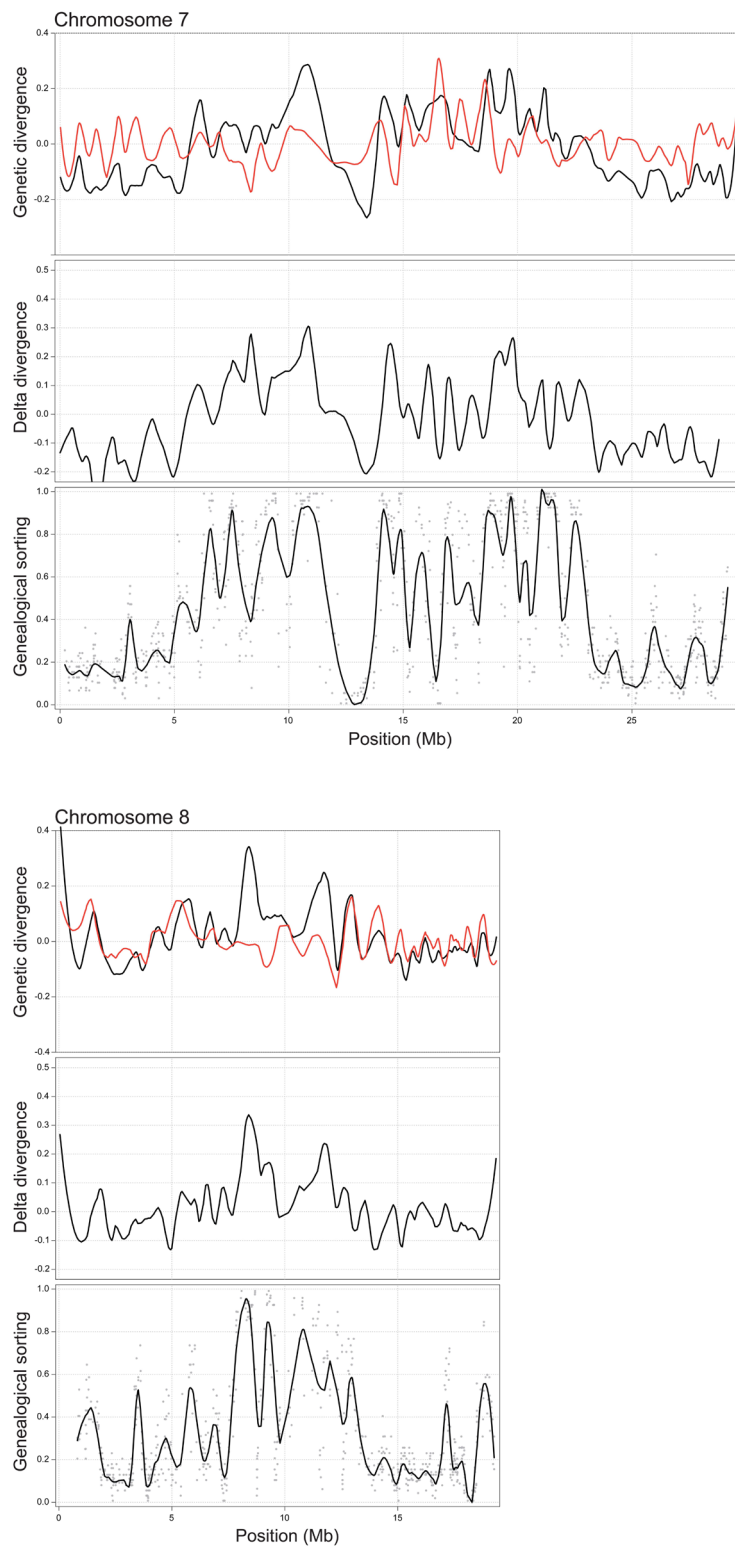


Figure S2 (chromosomes 9 and 10)

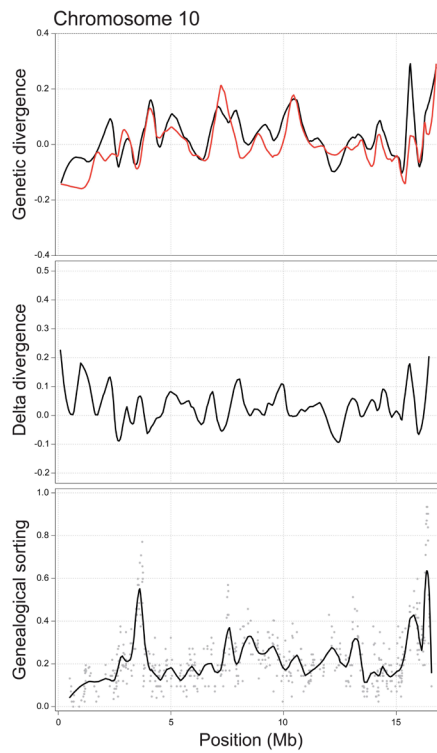
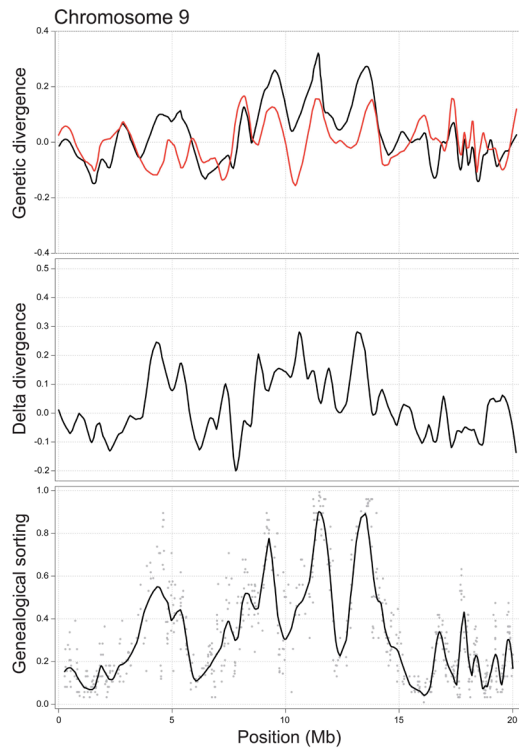


Figure S2 (chromosomes 11 and 12)

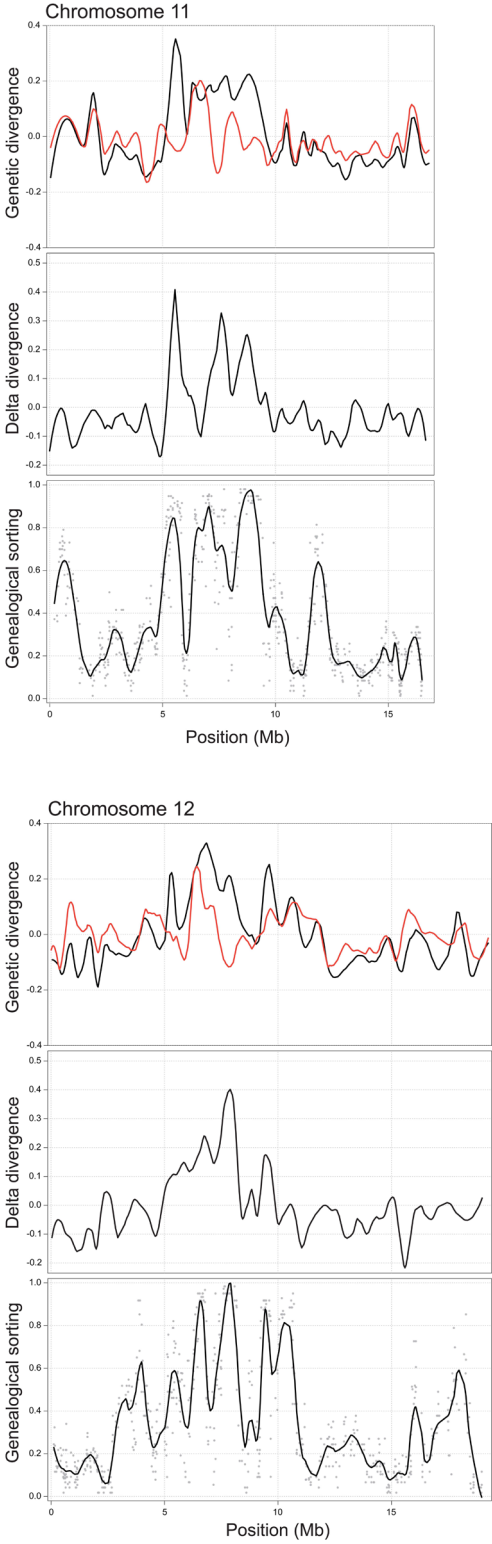


Figure S2 (chromosomes 13 and 14)

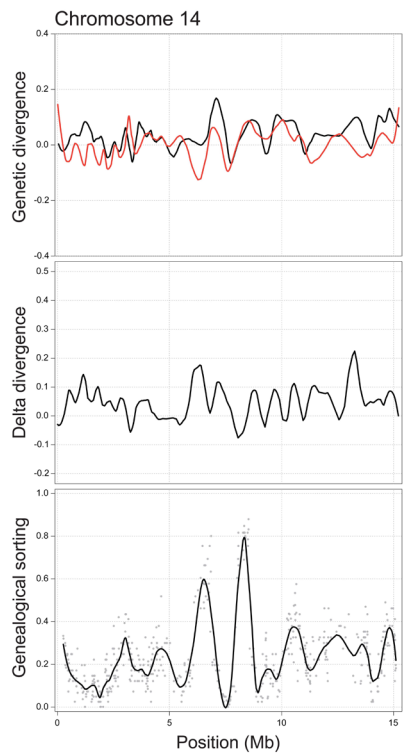
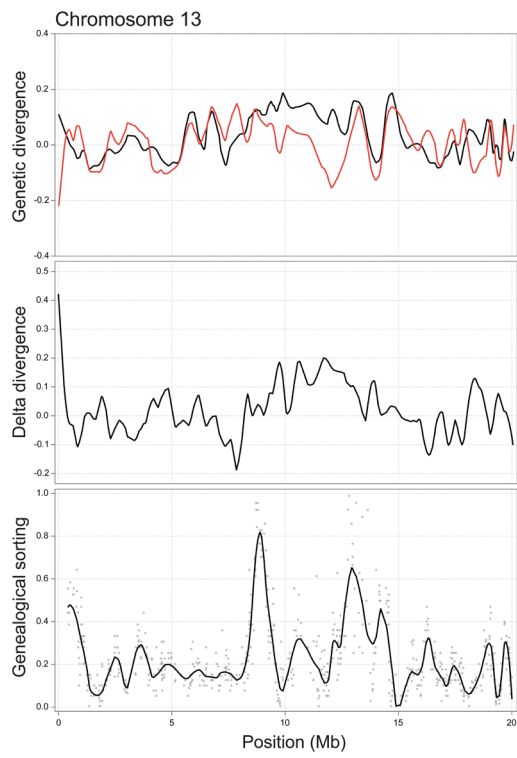


Figure S2 (chromosomes 15 and 16)

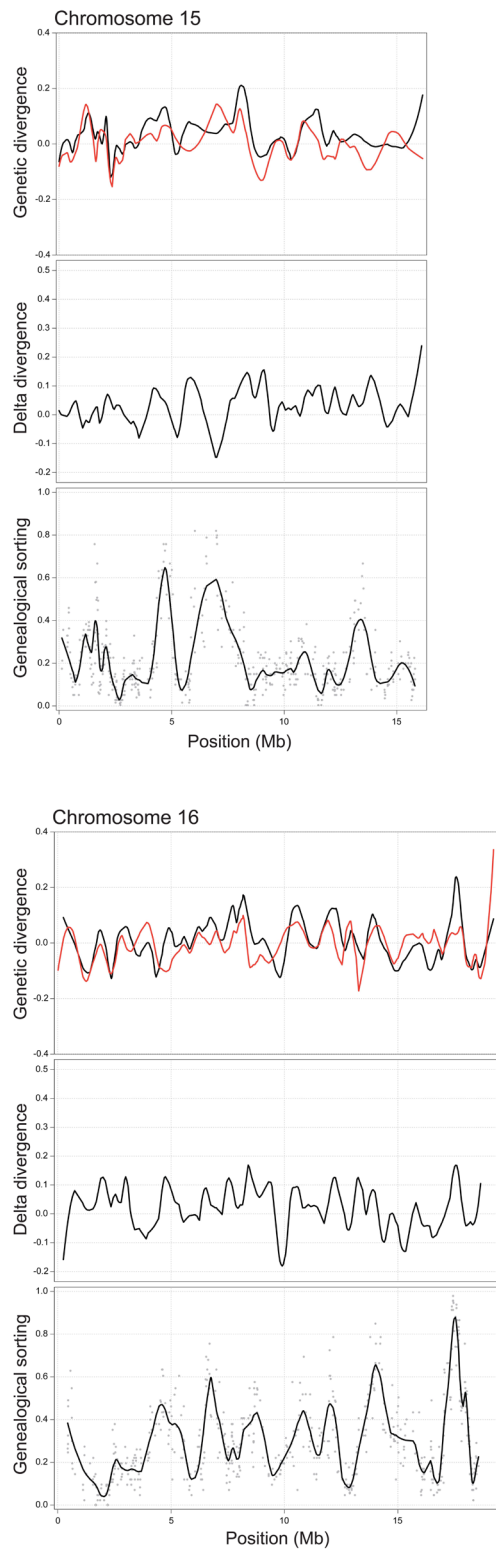


Figure S2 (chromosomes 17 and 18)

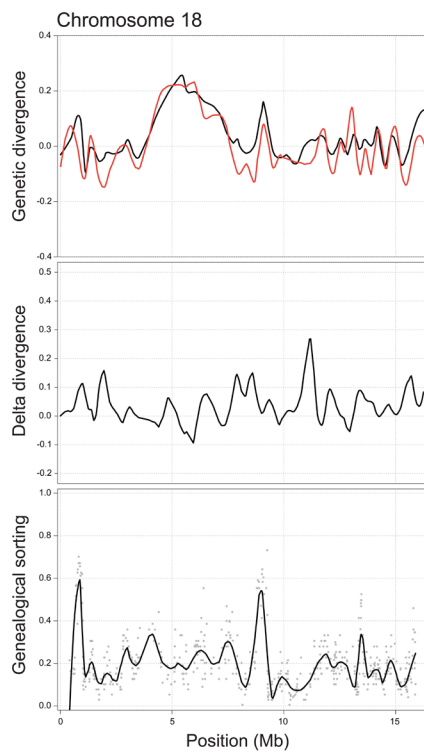
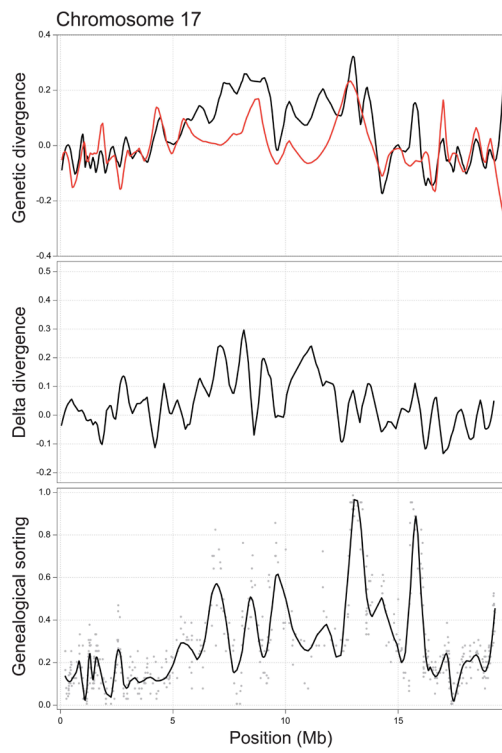
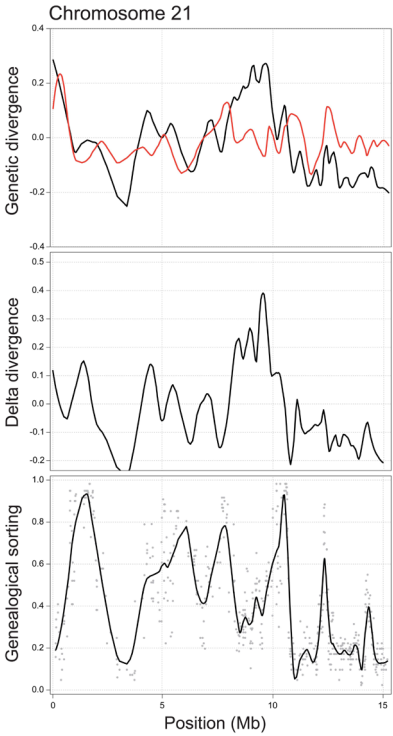
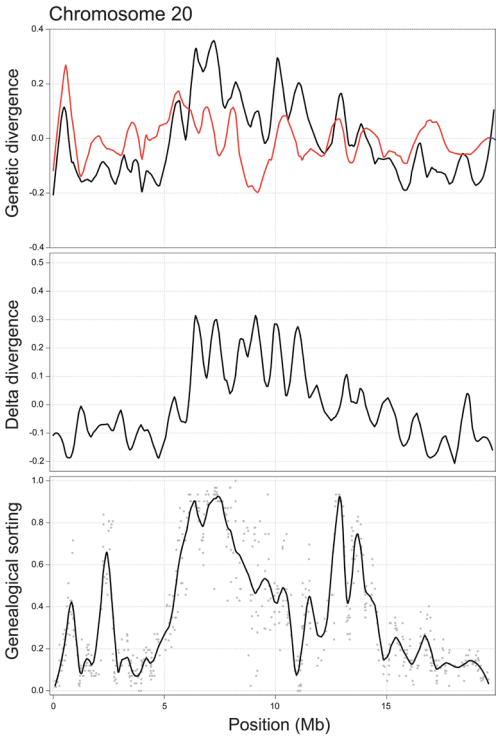


Figure S2 (chromosomes 20 and 21)



Supporting References

- Baird NA, Etter PD, Atwood TS *et al.* (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE*, **3**, e3376.
- Barrett SCH, Rogers SM, & Schluter D (2008) Natural selection on a major armor gene in threespine stickleback. *Science*, **322**, 255-257.
- Barrett RDH & Schluter D (2008) Adaptation from standing genetic variation. *Trends in Ecology and Evolution*, **23**, 38-44.
- Bell MA & Foster SA (1994) *The evolutionary biology of the threespine stickleback* (Oxford University, Oxford).
- Berner D, Adams DC, Grandchamp AC, & Hendry AP (2008) Natural selection drives patterns of lake-stream divergence in stickleback foraging morphology. *Journal of Evolutionary Biology*, **21**, 1653-1665.
- Berner D, Grandchamp A-C, & Hendry AP (2009) Variable progress toward ecological speciation in parapatry: stickleback across eight lake-stream transitions. *Evolution*, **63**, 1740-1753.
- Berner D, Stutz WE, & Bolnick DI (2010a) Foraging trait (co)variances in stickleback evolve deterministically and do not predict trajectories of adaptive diversification. *Evolution*, **64**, 2265-2277.
- Berner D, Roesti M, Hendry AP, & Salzburger W (2010b) Constraints on speciation suggested by comparing lake-stream stickleback divergence across two continents. *Molecular Ecology*, **19**, 4963-4978.
- Colosimo PF, Hosemann KE, Balabhadra S *et al.* (2005) Widespread parallel evolution in sticklebacks by repeated fixation of ectodysplasin alleles. *Science*, **307**, 1928-1933.
- Cummings MP, Neel MC, & Shaw KL (2008) A genealogical approach to quantifying lineage divergence. *Evolution*, **62**, 2411-2422.
- Deagle BE, Jones FC, Chan YF *et al.* (2012) Population genomics of parallel phenotypic evolution in stickleback across stream-lake ecological transitions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **279**, 1277-1286.
- Deen PM & Robben JH (2011) Succinate receptors in the kidney. *Journal of the American Society of Nephrology*, **22**, 1416-1422.
- DeFaveri J, Shikano T, Shimada Y, Goto A, & Merilä J (2011) Global analysis of genes involved in

- freshwater adaptation in threespine sticklebacks (*Gasterosteus aculeatus*). *Evolution*, **65**, 1800-1807.
- Dejima K, Murata D, Mizuguchi S *et al.* (2009) The ortholog of human solute carrier family 35 member B1 (UDP-galactose transporter-related protein 1) is involved in maintenance of ER homeostasis and essential for larval development in *Caenorhabditis elegans*. *The FASEB Journal*, **23**, 2215-2225.
- Domingues VS, Poh Y-P, Peterson BK *et al.* (2012) Evidence of adaptation from ancestral variation in young populations of beach mice. *Evolution*, **66**, 3209-3223.
- Evans DH, Piermarini PM, & Choe KP (2005) The multifunctional fish gill: dominant site of gas exchange, osmoregulation, acid-base regulation, and excretion of nitrogenous waste. *Physiological Reviews*, **85**, 97-177.
- Feder JL, Egan SP, & Nosil P (2012) The genomics of speciation-with-gene-flow. *Trends in Genetics*, **28**, 342-350.
- Felsenstein J (1984) Distance methods for inferring phylogenies: A justification. *Evolution*, **38**, 16-24.
- Greenwood AK, Jones FC, Chan YF *et al.* (2011) The genetic basis of divergent pigment patterns in juvenile threespine sticklebacks. *Heredity*, **107**, 155-166.
- Gross JB & Wilkens H (2013) Albinism in phylogenetically and geographically distinct populations of *Astyanax* cavefish arises through the same loss-of-function *Oca2* allele. *Heredity*, **111**, 122-130.
- Hermisson J & Pennings PS (2005) Soft sweeps: molecular population genetics of adaptation from standing genetic variation. *Genetics*, **169**, 2335-2352.
- Hoegg S, Brinkmann H, Taylor JS, & Meyer A (2004) Phylogenetic timing of the fish-specific genome duplication correlates with the diversification of teleost fish. *Journal of Molecular Evolution*, **59**, 190-203.
- Hohenlohe PA, Bassham S, Etter PD *et al.* (2010) Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLoS Genetics*, **6**, e1000862.
- Inokuchi M, Hiroi J, Watanabe S, Lee KM, & Kaneko T (2008) Gene expression and morphological localization of NHE3, NCC and NKCC1a in branchial mitochondria-rich cells of Mozambique tilapia (*Oreochromis mossambicus*) acclimated to a wide range of salinities. *Comparative Biochemistry and Physiology*, **151**, 151-158.
- Jakobsson S, Borg B, Haux C, & Hyllner SJ (1999) An 11-ketotestosterone induced kidney-secreted protein: the nest building glue from male three-spined stickleback, *Gasterosteus*

- aculeatus*. *Fish Physiology and Biochemistry*, **20**, 79-85.
- Kawahara R & Nishida M (2007) Extensive lineage-specific gene duplication and evolution of the spiggin multi-gene family in stickleback. *BMC Evolutionary Biology*, **7**, 209.
- Koga A, Inagaki H, Bessho Y, & Hori H (1995) Insertion of a novel transposable element in the thyrosinase gene is responsible for an albino mutation in the medaka fish *Oryzias latipes*. *Molecular Biology and Evolution*, **240**, 400-405.
- Jones FC, Chan YF, Schmutz J *et al.* (2012a) A genome-wide SNP genotyping array reveals patterns of global and repeated species-pair divergence in sticklebacks. *Current Biology*, **22**, 83-90.
- Jones FC, Grabherr MG, Chan YF *et al.* (2012b) The genomic basis of adaptive evolution in threespine sticklebacks. *Nature*, **484**, 55-61.
- Kot M (2001) *Elements of Mathematical Ecology* (Cambridge University Press, Cambridge, UK).
- Larsen PF, Nielsen EE, Koed A *et al.* (2008) Interspecific differences in expression of candidate genes for salinity tolerance in winter migrating anadromous brown trout (*Salmo trutta* L.). *BMC Genetics*, **9**, 12.
- Malinowska DH, Sherry AM, Tewari KP, & Cuppoletti J (2003) Gastric parietal cell secretory membrane contains PKA- and acid-activated Kir2.1 K⁺ channels. *American Journal of Physiology - Cell Physiology*, **286**, 495-506.
- Marchinko KB (2009) Predation's role in repeated phenotypic and genetic divergence of armor in threespine stickleback. *Evolution*, **63**, 127-138.
- Mateus CS, Stange M, Berner D *et al.* (2013) Strong genome-wide divergence between sympatric European river and brook lampreys. *Current Biology*, **23**, 649-650.
- McCairns RJS & Bernatchez L (2010) Adaptive divergence between freshwater and marine sticklebacks: insights into the role of phenotypic plasticity from an integrated analysis of candidate gene expression. *Evolution*, **64**, 1029-1047.
- McCormick SD (2001) Endocrine control of osmoregulation in teleost fish. *American Zoologist*, **41**, 781-794.
- Messer PW & Petrov DA (2013) Population genomics of rapid adaptation by soft selective sweeps. *Trends in Ecology and Evolution*, **28**, 659-669.
- Nadeau NJ, Whibley A, Jones RT *et al.* (2012) Genomic islands of divergence in hybridizing *Heliconius* butterflies identified by large-scale targeted sequencing. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **367**, 343-353.

- Nei M & Tajima F (1981) DNA polymorphism detectable by restriction endonucleases. *Genetics*, **97**, 145-163.
- Nosil P, Funk DJ, & Ortiz-Barrientos D (2009) Divergent selection and heterogeneous genomic divergence. *Molecular Ecology*, **18**, 375-402.
- Page-McCaw PS, Chung SC, Muto A *et al.* (2004) Retinal network adaptation to bright light requires tyrosinase. *Nature Neuroscience*, **7**, 1329-1336.
- Paradis E, Claude J, & Strimmer K (2004) Analyses of Phylogenetics and Evolution in R language. *Bioinformatics*, **20**, 289-290.
- Posada D (2008) jModelTest: Phylogenetic model averaging. *Molecular Biology and Evolution*, **25**, 1253-1256.
- Reilly BD, Cramp RL, Wilson JM, Campbell HA, & Franklin CE (2011) Branchial osmoregulation in the euryhaline bull shark, *Carcharhinus leucas*: a molecular analysis of ion transporters. *Journal of Experimental Biology*, **214**, 2883-2895.
- Reimchen TE (1992) Injuries on stickleback from attacks by a toothed predator (*Oncorhynchus*) and implication for the evolution of lateral plates. *Evolution*, **46**, 1224-1230.
- Reimchen TE (1994) Predators and morphological evolution in threespine stickleback. *The evolutionary biology of the threespine stickleback*, eds Bell MA & Foster SA (Oxford University, Oxford), pp 240-273.
- Renaut S, Nolte AW, Rogers SM, Derome N, & Bernatchez L (2011) SNP signatures of selection on standing genetic variation and their association with adaptive phenotypes along gradients of ecological speciation in lake whitefish species pairs (*Coregonus* spp.). *Molecular Ecology*, **20**, 545-559.
- Roesti M, Hendry AP, Salzburger W, & Berner D (2012a) Genome divergence during evolutionary diversification as revealed in replicate lake-stream stickleback population pairs. *Molecular Ecology*, **21**, 2852-2862.
- Roesti M, Salzburger W, & Berner D (2012b) Uninformative polymorphisms bias genome scans for signatures of selection. *BMC Evolutionary Biology*, **12**, 94.
- Roesti M, Moser D, & Berner D (2013) Recombination in the threespine stickleback genome – patterns and consequences. *Molecular Ecology*, **22**, 3014-3027.
- Salzburger W, Ewing GB, & von Haeseler A (2011) The performance of phylogenetic algorithms in estimating haplotype genealogies with migration. *Molecular Ecology*, **20**, 1952-1963.

- Schluter D & Conte GL (2009) Genetics and ecological speciation. *Proceedings of the National Academy of Sciences, USA*, **106**, 9955-9962.
- Scott GR, Rogers JT, Richards JG, Wood MC, & Schulte PM (2004) Intraspecific divergence of ionoregulatory physiology in the euryhaline teleost *Fundulus heteroclitus*: possible mechanisms of freshwater adaptation. *Journal of Experimental Biology*, **207**, 3399-3410.
- Shimada Y, Shikano T, & Merilä J (2011) A high incidence of selection on physiologically important genes in the three-spined stickleback, *Gasterosteus aculeatus*. *Molecular Biology and Evolution*, **28**, 181-193.
- Stephens M, Smith NJ, & Donnelly P (2001) A new statistical method for haplotype reconstruction from population data. *The American Journal of Human Genetics*, **68**, 978-989.
- Stephens M & Donnelly P (2003) A comparison of bayesian methods for haplotype reconstruction from population genotype data. *The American Journal of Human Genetics*, **73**, 1162-1169.
- Streisfeld MA, Young WN, & Sobel JM (2013) Divergent selection drives genetic differentiation in an R2R3-MYB transcription factor that contributes to incipient speciation in *Mimulus aurantiacus*. *PLoS Genetics*, **9**, e1003385.
- Swofford DL (2003) *PAUP*: Phylogenetic analysis using parsimony (*and other methods)* (Sinauer Associates, Sunderland).
- Team RDC (2013) *R: A language and environment for statistical computing* (R Foundation for Statistical Computing, Vienna, Austria).
- Tennessen JA, Akey JM (2011) Parallel adaptive divergence among geographically diverse human populations. *PLoS Genetics*, **7**, e1002127.
- Terai Y, Seehausen O, Sasaki T *et al.* (2006) Divergent selection on opsins drives incipient speciation in Lake Victoria cichlids. *PLoS Biology*, **4**, e433.
- van Rooijen E, Voest EE, Logister I *et al.* (2009) Zebrafish mutants in the von Hippel-Lindau tumor suppressor display a hypoxic response and recapitulate key aspects of Chuvash polycythemia. *Blood*, **113**, 6449-6460.
- Walker JA & Bell MA (2000) Net evolutionary trajectories of body shape evolution within a microgeographic radiation of threespine sticklebacks (*Gasterosteus aculeatus*). *Journal of Zoology*, **252**, 293-302.
- Wang X, Tan Y, Sievers Q *et al.* (2011) Thyroid hormone-response genes mediate otolith growth and development during flatfish metamorphosis. *Comparative Biochemistry and Physiology –*

Part A: Comparative Physiology, **158**, 163-168.

Weir BS & Cockerham CC (1984) Estimating F-statistics for the analysis of population-structure.

Evolution, **38**, 1358-1370.

Wootton RJ (1976) *The biology of the sticklebacks* (Academic, London).

Chapter 5

Genetic architecture of skeletal evolution in European
lake and stream stickleback



GENETIC ARCHITECTURE OF SKELETAL EVOLUTION IN EUROPEAN LAKE AND STREAM STICKLEBACK

Daniel Berner,^{1,2} Dario Moser,¹ Marius Roesti,¹ Heinz Buescher,¹ and Walter Salzburger¹

¹Zoological Institute, University of Basel, Vesalgasse, 1, CH-4051 Basel, Switzerland

²E-mail: daniel.berner@unibas.ch

Received October 4, 2013

Accepted February 11, 2014

Advances in genomic techniques are greatly facilitating the study of molecular signatures of selection in diverging natural populations. Connecting these signatures to phenotypes under selection remains challenging, but benefits from dissections of the genetic architecture of adaptive divergence. We here perform quantitative trait locus (QTL) mapping using 488 F₂ individuals and 2011 single nucleotide polymorphisms (SNPs) to explore the genetic architecture of skeletal divergence in a lake-stream stickleback system from Central Europe. We find QTLs for gill raker, snout, and head length, vertebral number, and the extent of lateral plating (plate number and height). Although two large-effect loci emerge, QTL effect sizes are generally small. Examining the neighborhood of the QTL-linked SNPs identifies several genes involved in bone formation, which emerge as strong candidate genes for skeletal evolution. Finally, we use SNP data from the natural source populations to demonstrate that some SNPs linked to QTLs in our cross also exhibit striking allele frequency differences in the wild, suggesting a causal role of these QTLs in adaptive population divergence. Our study paves the way for comparative analyses across other (lake-stream) stickleback populations, and for functional investigations of the candidate genes.

KEY WORDS: Effect size, *Gasterosteus aculeatus*, lateral plate, QTL mapping, RAD sequencing, vertebral number.

Exploring the genetic basis of adaptation promises to illuminate several long-standing issues in biological diversification. These include the number and genomic location of genetic changes underlying adaptation, their role in developmental pathways, their phenotypic effects and resulting ecological consequences, and their predictability (Orr 1998; Barton and Keightley 2002; Phillips 2005; Hoekstra and Coyne 2007; Mitchell-Olds et al. 2007; Wray 2007; Arendt and Reznick 2008; Stern and Orgogozo 2008; Mackay et al. 2009; Rockman 2011; Wake et al. 2011; Yeaman and Whitlock 2011). Currently, perhaps the most popular approach to investigating the genetics of adaptation is divergence mapping (Nielsen 2005; Storz 2005; Oleksyk et al. 2010). Here a large set of genome-wide molecular markers is screened for putative signatures of divergent selection between ecologically distinct populations. In well-developed empirical systems, this is proving a powerful method for the discovery of genomic regions or candidate genes involved in adaptive divergence (e.g., Akey

et al. 2002; Voight et al. 2006; Hohenlohe et al. 2010; Lawniczak et al. 2010; Jones et al. 2012; Nadeau et al. 2012; Roesti et al. 2012a, 2014; Mateus et al. 2013; Stölting et al. 2013). A shortcoming of divergence mapping, however, is that in general molecular signatures alone cannot tell us much about the traits actually targeted by selection (Mitchell-Olds et al. 2007; Stinchcombe and Hoekstra 2008; Storz and Wheat 2010). In the years to come, we can thus anticipate a surge of information about genome regions putatively influenced by divergent selection in many organisms, but knowledge about the phenotypes transferring selection to the molecules is likely to lag behind. Understanding the genetics of adaptation will thus benefit greatly from the combination of purely genomic investigations with extensive data on the genetic architecture of phenotypic divergence, as for instance obtained by quantitative trait locus (QTL) mapping.

In the present study, we report a QTL mapping experiment performed in a powerful system for studying adaptive divergence:

lake and stream populations of threespine stickleback fish (*Gasterosteus aculeatus*). Following the retreat of the last Pleistocene ice sheets, the colonization of freshwater by ancestral marine stickleback has resulted in the establishment of numerous populations occurring in adjacent lake and stream habitats (Reimchen et al. 1985; Lavin and McPhail 1993; Thompson et al. 1997; Hendry and Taylor 2004; Berner et al. 2008, 2010a; Aguirre 2009; Bolnick et al. 2009; Deagle et al. 2012; Moser et al. 2012; Lucek et al. 2013; Ravinet et al. 2013). Lake and stream stickleback are often ecologically divergent, with the most consistent difference concerning their foraging modes: lake stickleback partly or exclusively exploit pelagic food resources (zooplankton), whereas stream stickleback generally use benthic prey (macroinvertebrates; Gross and Anderson 1984; Berner et al. 2009; Kaeuffer et al. 2012; Moser et al. 2012; Ravinet et al. 2013). This divergence in foraging modes is associated with relatively consistent phenotypic differences in traits presumably important for prey capture and handling, such as overall body shape and gill raker structure (Reimchen et al. 1985; Lavin and McPhail 1993; Berner et al. 2008, 2009, 2010a; Kaeuffer et al. 2012; Lucek et al. 2013; Ravinet et al. 2013). The existence of replicate, ecologically and phenotypically divergent population pairings makes lake-stream stickleback an appealing system for the search of molecular signatures of divergent selection. Indeed, divergence mapping has already been performed in some lake-stream stickleback systems (Deagle et al. 2012; Roesti et al. 2012a).

By contrast, very little is known about the genetics of phenotypic divergence between lake and stream stickleback. Quantitative genetic (common-garden) experiments have demonstrated a genetic basis to divergence in some foraging traits (Lavin and McPhail 1993; Sharpe et al. 2008; Berner et al. 2011), but QTL dissections of the genetic architecture of phenotypic divergence have yet to be performed. We here take up this challenge by using QTL mapping to explore the genetic basis of divergence in skeletal features between lake and stream stickleback populations from Central Europe.

Materials and Methods

CROSS

Our study is based on an F_2 intercross population derived from a single *in vitro* cross of a male from Lake Constance (sampled at the ROM lake site described in Berner et al. 2010a) with a female from a stream draining into Lake Geneva (the CHE stream site in Berner et al. 2010a). The F_2 panel comprises 492 individuals (251 males, 237 females) selected haphazardly at one year of age from 35 separate F_1 crosses, each produced by a unique full-sib pairing. All details on crossing, husbandry, and handling are exactly as described in Roesti et al. (2013), a recombination study

based on a subset of 282 individuals from the full F_2 population used here for QTL mapping.

All fish were euthanized with an overdose of MS-222, photographed immediately as described in Berner et al. (2009), and stored in absolute EtOH. After six months of preservation, a fin clip was taken for genetic analysis and each individual was subjected to a digital X-ray scan of the whole body and a higher resolution scan of the head. This was performed by using a Faxitron Digital Specimen Radiography System LX-60 (tube voltage 35 kV, tube current 0.3 mA), including a reference size scale in all scans.

PHENOTYPING

Our study focuses on aspects of skeletal morphology, here defined broadly as bone traits. The first trait of interest was the length of the gill rakers (bony tubercles) located on the first branchial arch (Fig. 1). Gill rakers are important to foraging because they influence prey retention and handling performance (Gerking 1994; Sanderson et al. 2001). In particular, longer gill rakers generally promote foraging on small prey items (such as zooplankton), whereas shorter gill rakers are favored in fish foraging on larger prey (such as macroinvertebrates). Indeed, the natural source populations of our cross are highly divergent in this trait, with the lake population displaying 25% longer size-corrected gill rakers than the stream population (standardized mean difference: 0.99; see Fig. 2 in Berner et al. 2010a), and this divergence coincides with distinct foraging modes: Lake Constance stickleback forage pelagically on zooplankton (Lucek et al. 2012; Moser et al. 2012), whereas their conspecifics from the CHE stream site feed on larger benthic macroinvertebrates (Berner et al. 2010a). Given that such concurrent divergence in gill raker length and prey utilization has also been found in other (lake-stream) stickleback systems (Gross and Anderson 1984; McPhail 1984; Schluter and McPhail 1992; Bolnick 2004; Berner et al. 2008, 2010a,b; Matthews et al. 2010; Ravinet et al. 2013), and even in distantly related fish species (Kahilainen and Ostbye 2006; Pfaender et al. 2011), the divergence between ROM and CHE stickleback is very likely adaptive. We note that benthic versus pelagic resource specialization often coincides with additional divergence in gill raker number, but because the source populations are not divergent in this trait (Berner et al. 2010a), we did not include this trait in the current analysis. Gill raker length was measured on the left first branchial arch of the preserved specimens under a stereomicroscope fitted with an ocular micrometer at 50 \times magnification (precision: 0.01 mm). We measured and then averaged the length of the rakers two to five (counted from the joint with the dorsal arch bone, see Berner et al. 2008).

Next, we considered two aspects of head morphology: snout length and overall head length (Fig. 1). The pelagic ROM lake population displays lower values for both traits relative to the

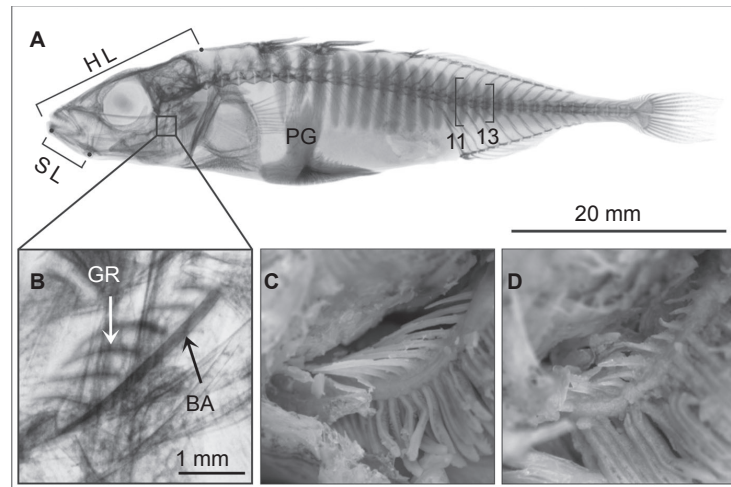


Figure 1. Traits subjected to QTL mapping in lake and stream stickleback. (A) X-ray scan of a partially plated stickleback visualizing snout length (SL), head length (HL), the vertebrae, and the lateral plates. Plate height was measured on the plates 11 and 13 posterior to the pelvic girdle (PG). (B) Detail of the head, showing the gill rakers (GR) on the first lower branchial arch (BA). The subpanels (C) and (D) display, on the same scale as (B), the gill rakers of two size-matched individuals from the upper and lower end of the gill raker length distribution.

CHE stream population foraging on benthic prey (Berner et al. 2010a), and similar foraging-related divergence is also seen in other stickleback systems (Caldecutt and Adams 1998; Albert et al. 2008). Moreover, head morphology typically shows sexual dimorphism (Caldecutt and Adams 1998; Kitano et al. 2007; Albert et al. 2008; Aguirre and Akinpelu 2010; Berner et al. 2010a, 2011; Ravinet et al. 2013). This dimorphism is possibly also related to differential foraging modes. The reason is that in stickleback, males tend to nest and provide parental care. The (presumably ancestral) necessity of males to forage on benthic resources during the breeding season while females can continue to exploit pelagic prey may have driven divergence in head structure between the sexes (Bentzen and McPhail 1984; Bentzen et al. 1984). Both snout length and head length were measured from the head X-ray scans. The former was defined as the distance from the joint to the tip of the lower jaw, the latter as the distance from the tip of the lower jaw to the dorsal posterior edge of the cranium.

The next trait quantified was vertebral number. Stickleback populations often differ in the number of vertebrae (Hagen and Gilbertson 1972; Moodie and Reimchen 1976; Reimchen et al. 1985). Although the functional basis of this variation remains poorly understood (but see Swain 1992), genetic analysis in stickleback may provide insights into vertebral diversification in other fish (Ward and Brainerd 2007; McDowall 2008) and vertebrates in general. As a first step, we thus produced whole-body X-ray scans of 14 specimens from each natural source population, counted all

vertebrae excluding the urostyle (Fig. 1), and tested for a population difference in mean count using 9999 random permutations of the trait over the populations (Manly 2007; all significance testing in this study was performed using analogous permutation procedures). This analysis made clear that ROM stickleback have a higher number of vertebrae than CHE fish (see Results). Following the same methods, we therefore quantified vertebral number for the full F_2 panel. Because of skeletal anomalies, 18 individuals could not be scored unambiguously, leaving 474 datapoints.

Finally, our phenotypic analysis included elements of lateral plating. Ancestral marine stickleback display a complete series of bony plates along their body, whereas the number of plates is typically greatly reduced in freshwater populations (Bell and Foster 1994). This difference is presumably attributable to differential exposure to predators (Hagen and Gilbertson 1972; Reimchen 1994; Bergstrom 2002; Leinonen et al. 2011a), although other ecological factors targeting plate number or other traits correlated with plate number due to pleiotropy or genetic linkage might influence plate evolution as well (e.g., Heuts 1947; Giles 1983; Barrett et al. 2009; Myhre and Klepaker 2009; Leinonen et al. 2011b; Roesti et al. 2014). Interestingly, Lake Constance stickleback are a rare example of a freshwater population almost completely fixed for the fully plated phenotype (Berner et al. 2010a; Moser et al. 2012). Because plating is reduced in several tributary streams to Lake Constance, the persistence of full plating in the lake likely reflects an adaptation to high predator exposure associated with a pelagic life style. By contrast, CHE stickleback represent a

typical low-plated freshwater population (Berner et al. 2010a), thus providing the opportunity to map variation in lateral plating in the F₂ cross. Consistent with previous work (Berner et al. 2010a; Moser et al. 2012), we scored the extent of lateral plating using three discrete phenotypic classes (low, partially, and fully plated).

Previous mapping efforts and subsequent functional analysis in Pacific marine and freshwater stickleback have already identified the *Ectodysplasin* (*EDA*) gene as a major driver of evolutionary shifts in the extent of lateral plating (Colosimo et al. 2004, 2005; Cresko et al. 2004; Baird et al. 2008). Moreover, targeted sequencing of the entire *EDA* coding region revealed distinct haplotypes in fully plated Lake Constance versus low-plated CHE stream fish (Berner et al. 2010a). Our primary objective in mapping plate morph was therefore to assess if genomic regions beyond *EDA* contribute to plating divergence between these European freshwater populations. To this end, we additionally counted the total number of lateral plates posterior to the pelvic girdle (including the plates forming the caudal keel) across both body sides in the subset of F₂ individuals genotyped unambiguously as heterozygotes at our single SNP marker located within *EDA*. Focusing on this particular subset ($N = 209$) allowed us to screen for loci influencing the extent of plating while controlling rigorously for the effect of the known major locus.

As an alternative to reducing the extent of lateral plating via a reduction in the number of plates (see above), stickleback sometimes appear to evolve shallower plates (Leinonen et al. 2012). Although differences in plate size between ROM and CHE stickleback could not be examined adequately because the latter are low-plated, a preliminary inspection of the F₂ population indicated substantial variation in plate height. We therefore measured the maximal height of the plates 11 and 13, as counted from the pelvic girdle, perpendicular to the anterior-posterior axis on the left body side (Fig. 1). Measurements were taken with a digital caliper (precision: 0.01 mm) handled under a stereomicroscope at 10–30 \times magnification. As plate height could only be quantified in the fully and most of the partially plated individuals, sample size was 358 and 342 for plate 11 and 13.

All metric (length) traits considered in our study scaled strongly with overall body size (Pearson's r : 0.49–0.84), whereas the meristic (count) traits did not (vertebral number: $r = 0.042$; lateral plate number: $r = 0.045$). Prior to QTL mapping, we therefore subjected the former traits to size correction by regressing each trait separately against body size, and treating the residuals as size-independent variables (Reist 1985; Berner 2011). These variables were shifted back into the original measurement range by adding the trait value predicted by the regression at mean body size across all individuals. To obtain a robust size metric for these procedures, we used *tpsDig* (Rohlf 2001) to digitize 16 landmarks as described in Berner et al. (2010a) on the digital photographs

of all individuals, and extracted geometric morphometric centroid size using *TpsRelw* (Rohlf 2001).

Finally, we assessed measurement precision for all traits by remeasuring 30 haphazardly selected individuals on a second occasion, and calculating the repeatability (Lessells and Boag 1987). Repeatability was consistently very high, ranging from 0.96 (plate number) to 1 (vertebral number, plate morph). The complete phenotype matrix used for mapping in R/qtl is available on the Dryad Digital Repository (doi:10.5061/dryad.b2534).

MARKER GENERATION

As markers for mapping, we used single nucleotide polymorphisms (SNPs) discovered by RAD sequencing (Baird et al. 2008). In brief, this involved DNA restriction with the *SbfI* enzyme, and sequencing pools of 62 barcoded individuals in eight lanes with 100 cycles on an Illumina HiSeq 2000 instrument. RAD library preparation and the bioinformatics pipeline used for SNP discovery and genotyping were exactly as described in Roesti et al. (2013). From the 2165 markers thus obtained, we excluded 154 to avoid *SbfI* restriction sites being represented by more than one SNP. We also discarded four individuals exhibiting more than 10% missing genotypes across all markers. Our final mapping dataset thus comprised 488 F₂ individuals (recall that sample size was lower for some traits) and 2011 SNPs. The genotype matrix used for mapping is available on the Dryad Digital Repository (doi:10.5061/dryad.b2534).

QTL MAPPING

All SNPs were ordered physically according to the stickleback genome reassembly performed in Roesti et al. (2013; available at <http://datadryad.org/resource/doi:10.5061/dryad.846nj.2>), resulting in 61–152 markers per chromosome and a median marker spacing of 118 kb. Next, we excluded 111 individuals with relatively low genotyping quality, as judged by clearly inflated genome-wide crossover count (we here used 2.5 times the median autosomal crossover count across all 488 individuals as threshold). The remaining 377 individuals were used to estimate the genetic map in R/qtl (Broman and Sen 2009), applying the Kosambi map function (assuming crossover interference). The resulting genetic map proved highly consistent with that provided in Roesti et al. (2013) based on fewer individuals but with genotype errors corrected manually, and was used to specify the genetic marker distances for QTL mapping. Mapping with genetic distances estimated by using the full F₂ panel produced very similar results (details not presented).

All phenotypes were subjected to single-QTL interval mapping in R/qtl using the extended Haley–Knott regression method (Broman and Sen 2009) and the full F₂ panel. Head length was mapped both with and without snout length entered as covariate, as our head length measure included the snout tip. We present

the former analysis only, noting that both approaches produced quantitatively very similar results. QTL significance was established based on the distribution of genome-wide maximum LOD (logarithm of the odds ratio for QTL likelihood) scores across 1000 random permutations of the phenotype data over the genotype data (Broman and Sen 2009). In the Results, we present only QTLs significant at the 0.05 level, but additional loci are considered in the Discussion, and a table including suggestive QTLs ($0.05 \leq P < 0.1$) is provided as Table S1. QTL effect sizes were quantified both as the percentage of the total phenotypic variance in the F_2 cross explained by the QTLs (percent variance explained, PVE), and as their allelic substitution effect (i.e., the phenotypic difference between the two homozygous genotype classes). We present the latter both in the traits' original measurement scale (millimeter for all length traits), and standardized by the average standard deviation within the homozygous genotype classes. All statistics and plotting were carried out with the R language (R Development Core Team 2013).

EXPLORING QTLs

Following QTL detection, we retrieved from the Ensembl Genome Browser all genes located in the physical window spanned by the two SNPs flanking the marker displaying the LOD peak (this interval usually coincided with the 1.5 LOD support interval). We then scanned these gene lists for strong causative candidates, as judged by information on protein function in vertebrate model organisms (chicken, mice, rats, humans) compiled in the UniProt database (The UniProt Consortium 2013).

In addition, the availability of RAD sequences generated previously for the ROM and CHE population allowed us to inspect the magnitude and direction of allele frequency shifts in the wild at QTLs discovered in the F_2 population. Although this type of follow-up analysis has, to our knowledge, not previously been carried out, it promised stronger QTL inference because a genotype–phenotype association shared between a cross and its natural source populations suggests that the focal QTL is effectively contributing to divergence in the wild, as opposed to being specific to the cross. As a caveat, we note that this approach assumes that the tight linkage detected between marker and QTL alleles in the cross also persists in the wild. Specifically, we here capitalized on RAD sequence data from 27 individuals sampled from each source population. Details on the wet laboratory protocol, the analysis pipeline, and access to the sequence data are provided in Roesti et al. (2012b; this reference describes data generation for the ROM population only; the CHE dataset has not previously been analyzed but was generated in exactly the same way). Because the RAD libraries of both the cross and the natural populations were generated using the *SbfI* restriction enzyme, all RAD loci of interest were shared among the two

datasets. However, the latter libraries were Illumina-sequenced to 76 bases as opposed to 100 bases for the cross, thus precluding the examination of allele frequencies in the natural populations at QTL-linked SNPs located distal to the restriction site. We further ignored SNPs linked to lateral plate height QTLs because we here lacked information on the direction and magnitude of divergence between the natural populations (see above).

For those SNPs represented in both the cross and the natural population datasets (four SNPs in total), we first determined from which population each of the two alleles present in the cross originated. This assignment was unambiguous because our study considered only markers homozygous within each grandparent (Roesti et al. 2013). Next, we arbitrarily converted the two alleles to integers (0, 1) and tested for frequency shifts by random permutation, using the difference in the population means as test statistic. While providing a formal test for population divergence at QTL-linked SNPs, this approach yielded no information regarding the potential cause of divergence at these SNPs. To gain insights into the latter, we performed a second analysis comparing the SNP allele frequency shifts between the natural populations to the magnitude of genome-wide baseline divergence between the populations. The rationale was that an allele frequency shift clearly exceeding baseline divergence—reflecting the magnitude of differentiation by drift—offers evidence for divergent selection having acted in the close neighborhood of the QTL-linked marker. We recognize the possibility, however, that selection may not have targeted the detected QTL itself, but a nearby locus unrelated to the mapped phenotype.

We thus translated allele frequency differences at the QTL-linked SNPs to F_{ST} (Nei and Tajima 1981, eq. 7), and estimated the confidence interval for F_{ST} as the 95 percentile of the distribution produced by bootstrap resampling the observed alleles 10,000 times within each population (Manly 2007). This confidence interval was then evaluated against the magnitude of baseline differentiation between the ROM and CHE population samples, defining baseline differentiation as genome-wide median F_{ST} (Roesti et al. 2012a). Following Roesti et al. (2012a,b), the estimation of baseline differentiation ignored SNPs with a minor allele frequency <0.25 to avoid polymorphisms with low information content, and for RAD sites harboring multiple polymorphisms used only the one SNP yielding the highest F_{ST} value. Baseline differentiation thus calculated was 0.37 across 5429 informative SNPs.

Results

GILL RAKER LENGTH

We found two significant QTLs influencing gill raker length (Table 1; an additional suggestive QTL is described in Table S1;

Table 1. Characterization of the QTLs for skeletal divergence between lake and stream stickleback.

Trait	Marker	Chromosome	Position (bp)	LOD	<i>P</i>	PVE	HSE	Direction	Candidate gene
Gill raker length	chrVI_12733534	6	13,735,445	4.52	0.027	6.5	0.19 (0.73)	L*	
	chrIV_570692	4	570,692	4.60	0.019	4.3	0.08 (0.33)	S	<i>BAPX1</i> (15)
Snout length	chrXIX_19432535	19	69,077 (contig 1730)	45.13	0.001	42.9	0.38 (1.87)	M	
Head length	chrXIX_19432535	19	69,077 (contig 1730)	5.70	0.002	7.7	0.33 (0.58)	M	
	chrUn_11709633	5	464,792	5.57	0.004	6.9	0.45 (0.72)	S*	
Vertebral number	chrXV_11777081	15	11,777,081	5.01	0.015	4.4	0.38 (0.58)	S*	
	chrXIV_6849438	14	6,849,438	4.55	0.031	3.2	0.24 (0.37)	S*	
	chrXXI_2306628	21	4,955,041	7.82	0.001	9.2	0.43 (0.85)	L*	<i>COL11A1</i> (6)
	chrXVII_1670571	17	1,670,571	7.64	0.001	6.4	0.44 (0.78)	L*	<i>ASPEN, OGN</i> (11)
Plate morph	chrIV_12797213	4	12,797,213	155.45	0.001	76.0	–	L*	<i>EDA</i> (15)
Plate 11 height	chrXI_10140558	11	10,140,558	9.69	0.001	12.7	0.79 (1.04)	S	<i>AXIN2</i> (36)
	chrXI_6239999	11	6,239,999	8.94	0.001	12.0	0.71 (0.97)	S	<i>PHOSPHO1</i> (36)
Plate 13 height	chrIV_4185607	4	4,185,607	4.54	0.041	5.5	0.47 (0.6)	L	
	chrIV_6474941	4	6,474,941	8.20	0.001	11.8	0.58 (1.07)	L	
	chrXI_10140558	11	10,140,558	5.96	0.002	9.4	0.6 (0.8)	S	<i>AXIN2</i> (36)
	chrXI_6239999	11	6,239,999	5.79	0.002	8.4	0.47 (0.69)	S	<i>PHOSPHO1</i> (36)
	chrIX_9659641	9	12,543,749	4.99	0.012	7.7	0.53 (0.71)	L	

The marker names specify genomic locations (chromosome and base pairs) according to the Broad S1 genome assembly, whereas the chromosome numbers and positions given in separate columns refer to the improved assembly (Roesti et al. 2013). The position of the marker on chromosome 19 (sex chromosome) is unclear (it proved linked relatively loosely to the other markers within the nonrecombining domain of this chromosome), hence we provide the position within its contig. Effect sizes are expressed as percent variance explained (PVE), and as homozygous substitution effect (HSE; in measurement unit, and standardized in parentheses). HSE is not given for plate morph, as this trait has an ordinal scale, and effect sizes for the two plate height QTLs on chromosome 11 are probably inflated because of linkage. The QTLs are ordered by PVE within each trait. The Direction column indicates whether ROM lake (L), CHE stream (S), or male (Y-linked; M) alleles cause higher trait values, and asterisks indicate allelic effects in the direction expected from the divergence between the natural populations (note that this could not be determined for the plate height QTL). The last column lists candidate genes found in the marker intervals around the QTL SNPs, with the numbers in parentheses indicating the total number of genes in each interval (including predicted genes). This table reports only QTLs reaching $P < 0.05$; additional suggestive loci are presented in Table S1.

genome-wide LOD profiles for all traits are presented as Fig. S1). Both showed a modest effect size. The SNP associated with the QTL exhibiting the greater effect size (located on chromosome 6) produced a phenotypic shift in the predicted direction (longer gill rakers associated with the ROM lake allele), while the other one did not. Only the chromosome 6 marker could be analyzed for allele frequency shifts in the natural populations, revealing almost complete fixation of the expected SNP allele within the ROM and CHE sample (permutation $P = 0.0001$). This frequency shift ($F_{ST} = 0.77$, lower and upper 95% confidence limits: 0.65, 0.92) was much stronger than expected from the populations' baseline divergence.

HEAD MORPHOLOGY

The analysis of snout length detected a single large-effect QTL only (43 PVE; Table 1, Fig. S1). This QTL mapped to the domain on the sex chromosome (19) where the X and Y gametologs do not recombine (Roesti et al. 2013). Males of the F₂ population further displayed strikingly longer snouts than females (Fig. 2). Together, these observations indicated very strong sex-linked control of snout length. Indeed, mapping sex as a binary trait produced a single significant QTL (LOD = 380) coinciding exactly with the snout length QTL, whereas mapping snout length separately within each sex produced no QTL (details not presented).

Some sex-linked control was observed for overall head length as well, as the snout length QTL was also the strongest QTL affecting head length (Tables 1, S1, Fig. S1; recall that head length was mapped with snout length as covariate, so this finding is not a methodological artifact). Three additional minor head length QTLs were detected on the autosomes, all of them (and also the suggestive locus) exhibiting an effect in the direction predicted from previous phenotypic work (a larger head associated with the CHE stream alleles). Allele frequencies in the natural populations could be inspected for the marker linked to the QTL on chromosome 15 only, which again revealed a shift in the predicted direction ($P = 0.0001$; $F_{ST} = 0.51$, CLs: 0.35, 0.68) and exceeding the baseline level.

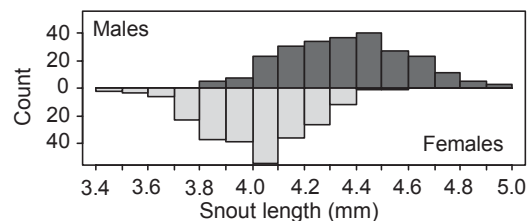


Figure 2. Distribution of size-adjusted snout length in male (dark gray, pointing upward) and female (light gray, pointing downward) stickleback from the F₂ panel.

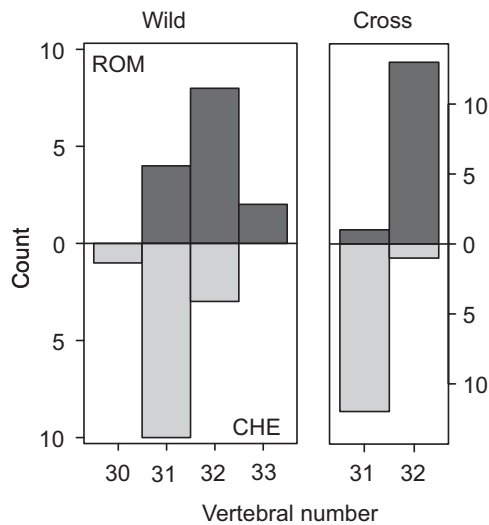


Figure 3. Vertebral number in ROM lake and CHE stream stickleback. The left panel shows vertebral count histograms based on a sample ($N = 14$) from the lake population (dark gray, pointing upward) and the stream population (light gray, pointing downward). The right panel displays the distribution of vertebral number in individuals from the F_2 cross concurrently homozygous for either the lake alleles (dark gray; $N = 14$) or the stream alleles (light gray; $N = 13$) at the two QTLs identified on chromosomes 17 and 21.

VERTEBRAL NUMBER

Most stickleback from the ROM lake sample displayed 32 vertebrae, as opposed to 31 vertebrae predominating in CHE stream fish ($P = 0.0085$; Fig. 3). In the cross, vertebral number mapped to two QTLs, with their effects being in the expected direction (ROM lake alleles associated with greater vertebral number; the same is true for the suggestive loci; Tables 1, S1, Fig. S1). Despite moderate effect sizes of the two QTLs when estimated separately (9.2 and 6.4 PVE), their joint effect was striking: individuals homozygous for the ROM lake or CHE stream alleles at both QTL-linked SNPs simultaneously exhibited almost consistently 32 versus 31 vertebrae (Fig. 3). Comparing a subset of F_2 individuals with 31 and 32 vertebrae ($N = 30$ each) showed unambiguously that the variation was in the number of caudal as opposed to abdominal vertebrae (details not presented). Moreover, testing for a difference in mean body size between individuals with 31 versus 32 vertebrae (together accounting for 97% of all F_2 individuals) revealed clearly that vertebral number and body size were unrelated ($P = 0.36$, standardized mean difference in size between the two groups: 0.06; visualized in Fig. S2). Allele frequency shifts could be examined for the marker linked to the QTL on chromosome 21 only, revealing divergence in the expected direction ($P = 0.0001$;

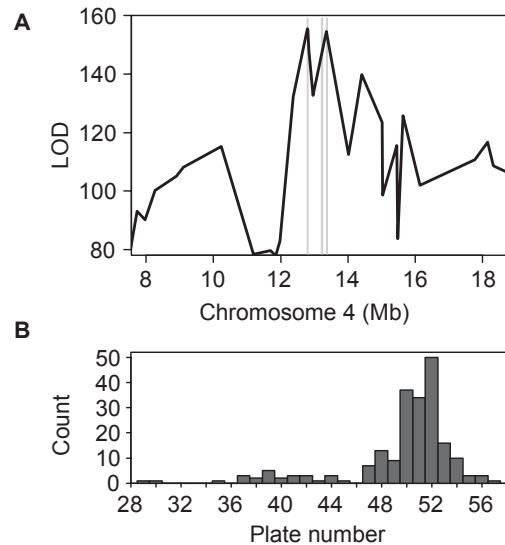


Figure 4. (A) LOD profile for the segment on chromosome 4 containing the detected QTL for lateral plate morph. The gray vertical lines indicate, from left to right, the position of *EDA* and the two candidate genes *PDLIM7* and *ANXA6*. The two LOD peaks (>150) are separated by two markers. (B) Distribution of the total number of lateral plates posterior to the pelvic girdle in F_2 individuals heterozygous at the *EDA* marker.

$F_{ST} = 0.41$, CLs: 0.15, 0.71), but not stronger than expected from the baseline.

LATERAL PLATING

Mapping lateral plate morph detected a locus of large effect (76 PVE) on chromosome 4, located precisely in the *EDA* region (LOD = 155.4 for the SNP 3 kb from the start position of *EDA*, and LOD = 155.2 for the SNP within *EDA*; Table 1, Fig. 4A). An almost equally strong marker-phenotype association (LOD = 154.6), however, occurred at 13.35 Mb. Inspecting the genotype frequencies at the two SNPs separating these high-LOD regions indicated that the drop in the strength of marker-phenotype association was not due to low genotyping quality (details not presented). At the *EDA* SNP, the natural populations were relatively close to fixation for the expected alternative alleles ($P = 0.0001$; $F_{ST} = 0.60$, CLs: 0.42–0.79), a shift clearly exceeding the baseline divergence level. No additional plate morph QTL was found (Fig. S1).

Lateral plate number was variable among the F_2 individuals heterozygous at *EDA*, but never lower than 29 (Fig. 4B). Hence, *EDA* heterozygotes in our cross always classified as either partially or fully plated. We found no significant QTL for plate

number when only considering *EDA* heterozygotes (Fig. S1; a single suggestive QTL is described in Table S1).

The height of lateral plate 11 mapped to two QTLs (Tables 1, S1, Fig. S1). The locus on chromosome 11 displayed a substantial effect size (12.7 PVE), with the ROM lake allele associated with shorter plates. However, this effect size was probably slightly inflated, as inspecting the LOD profile along chromosome 11 revealed the presence of an additional, nearly equally strong QTL 4 Mb (9.4 cM) away (LOD = 8.94; this second QTL is included in Table 1). Lacking information on plate height divergence between the natural populations, we did not investigate population-level shifts in allele frequencies.

The genetic architecture of lateral plate 13 height overlapped partly with that of plate 11 (Table 1, Fig. S1), which is not surprising, given the relatively strong phenotypic correlation of the two plate height traits within the F_2 population ($r = 0.84$). Specifically, we detected exactly the same two QTLs located on chromosome 11. However, the strongest effect (11.8 PVE) was now seen in a QTL on chromosome 4. This locus was 2.3 Mb away from the QTL on the same chromosome driving plate 11 height (Fig. S1), and thus perhaps represents a distinct locus, although this cannot be determined with confidence. In addition, plate 13 height was influenced by a QTL on chromosome 9.

Discussion

We have used an F_2 intercross to investigate the genetic architecture of divergence in skeletal traits between lake and stream stickleback. A first suite of traits considered included gill raker, snout, and head length, traits believed to mediate trophic specialization. In particular, gill raker length displays a highly predictable association with prey use in stickleback and other fish species. To our knowledge, variation in gill raker length has not previously been mapped in any species, but common-garden experiments in stickleback indicated a heritable basis to the phenotypic divergence between benthic and pelagic populations (McPhail 1984; Day et al. 1994; Wund et al. 2008). In line with these quantitative genetic observations, our study discovered QTLs for gill raker length. Despite a mutation screen in zebrafish, implicating the *Ectodysplasin (EDA)* signaling pathway in gill raker formation (Harris et al. 2008), the detected QTLs showed no obvious relationship to that pathway. However, screening the marker interval around the QTL on chromosome 4 suggested *BAPX1* as a strong candidate gene, given that *BAPX1* is crucial to the formation of the first branchial arch in zebrafish (Miller et al. 2003).

We also found several autosomal QTLs explaining variation in head length. The effect sizes were consistently in the direction expected from the phenotypic divergence between the source populations (Berner et al. 2010a) and other benthic-pelagic stickleback systems (Caldecutt and Adams 1998; Albert et al. 2008).

Nevertheless, in accordance with greater male than female overall head length found in many stickleback studies (Caldecutt and Adams 1998; Kitano et al. 2007; Albert et al. 2008; Aguirre and Akinpelu 2010; Berner et al. 2010a, 2011; Ravinet et al. 2013), the strongest head length QTL was sex-linked and also turned out to be the only (large-effect) locus driving snout length. Hence, while contrasting foraging habitats likely drive the evolution of stickleback head morphology among populations, the footprint of sex-specific selection is much stronger.

VERTEBRAL NUMBER

The source populations of the cross showed clear divergence in the number of vertebrae, with a higher average count in lake than stream fish. This trend has also been found in other studies comparing lake and stream stickleback (Hagen and Gilbertson 1972; Reimchen et al. 1985) and thus likely represents an adaptive response to divergent selection on locomotion (Swain 1992). Although further functional evidence is needed, our finding that vertebral number is genetically unrelated to body size (as also found in a different stickleback system; Alho et al. 2011) indicates that population divergence in the number of vertebrae is unlikely to reflect a correlated response to selection on size (note that the ROM lake and CHE stream populations differ in size; Moser et al. 2012).

In our cross, vertebral number mapped to two QTLs. These loci explained a moderate proportion of the total phenotypic variance when considered in isolation, but in combination had a high explanatory power: their joint homozygous substitution accounted for an approximate shift of one vertebra, roughly the magnitude of divergence between the natural populations. To our knowledge, vertebral number has previously been mapped only in two fish species (medaka and trout) and in pigs. The former studies detected QTLs but were performed with a marker resolution too low to allow candidate gene identification (Nichols et al. 2004; Kimura et al. 2012). QTLs also emerged in pigs, where fine mapping produced strong candidate genes (*NR6A1*, *VRTN*, *PROX2*, *FOS*; Mikawa et al. 2007, 2011; Ren et al. 2012). The stickleback homologues of these genes, however, are not located on the chromosomes 17 and 21 where we found QTLs for vertebral number. Instead, screening the target marker interval on chromosome 17 identified *OGN* (Madisen et al. 1990) and *ASPN* as candidate genes. Both genes are involved in bone formation. In particular, *ASPN* regulates osteoblast collagen mineralization *in vitro* (Kalamajski et al. 2009) and is implicated in human degenerative diseases of skeletal joint regions, including intervertebral disks (Kizawa et al. 2005; Song et al. 2008). Also, a strong candidate gene (*COL11A1*) emerged in the focal chromosome 21 segment. Mutations in *COL11A1* cause skeletal disorders, including the malformation of vertebrae (Li et al. 1995; Tompson et al. 2010; Koyama et al. 2012).

LATERAL PLATING

At first glance, our mapping of lateral plate morph produced an expected result: the LOD maximum mapped to the *EDA* gene, and the corresponding SNP explained 76% of the total variance, a value very similar to that reported in Colosimo et al. (2004) for the plate morph QTL on chromosome 4 (77.6 PVE). Interestingly, however, a nearly equally strong LOD score emerged at 13.35 Mb (roughly 0.5 Mb from *EDA*), a region identified as high differentiation outlier in a divergence mapping study using fully plated marine and low-plated freshwater stickleback from Alaska (Hohenlohe et al. 2010). Our marker interval in that region contains two strong candidate genes, *PDLIM7* and *ANXA6* (see also Hohenlohe et al. 2010). *PDLIM7* has been shown to initiate bone formation de novo, and also to interact with bone morphogenetic protein (*BMP*) signaling (Boden et al. 1998; Liu et al. 2002). Similarly, *ANXA6* plays a critical role during the calcification of skeletal tissue (Kirsch et al. 2000; Wang and Kirsch 2002; Thouverey et al. 2009). It is thus possible that the strong effect seen at the *EDA* marker in our cross captures variation in lateral plating driven by polymorphism in one or more additional genes in its close neighborhood. Unfortunately, the paucity of crossovers between the SNPs at these QTLs in our cross precludes disentangling their relative effect sizes.

Outside chromosome 4, we found no QTL substantially influencing plate morph or number. This is surprising, given that such QTLs were discovered previously on chromosomes 7, 10, and 21 (Colosimo et al. 2004; but see Baird et al. 2008). This difference in genetic architecture likely explains why *EDA* heterozygotes in our cross were never low-plated, although low-plated heterozygotes occurred in the mapping panel studied by Colosimo et al. (2004).

As a complementary route to the adaptive reduction in lateral plating, stickleback might evolve shallower plates (Leinonen et al. 2012). Mapping the height of the plates 11 and 13 posterior to the pelvic girdle, we found QTLs on chromosomes 4, 9, and 11. These results differ from the previous report of plate height QTLs on the chromosomes 4 (at around 2 Mb, hence in a different region than in our cross), 7, and 20 (Colosimo et al. 2004). However, that study measured plates immediately adjacent to the pelvic girdle. Combined with our observation that the relative influence of the QTLs on chromosomes 4 and 11 on the height of the plates 11 and 13 was inverted, and that the QTL on chromosome 9 (and the additional suggestive QTLs) influenced one of the plates only, we conclude that plate height has a fairly complex genetic architecture, with several loci acting relatively locally.

Examining the plate height QTL regions produced strong candidate genes. Notably, the marker interval around the highest LOD peak observed (chromosome 11) included *AXIN2*. Loss of function mutations in *AXIN2* lead to ectodermal dysplasia in humans (Lammi et al. 2004; Mostowska et al. 2006; Callahan et al.

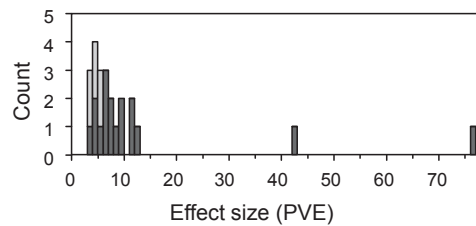


Figure 5. Distribution of effect sizes (in percent variance explained) across all QTLs (including in light gray the suggestive ones with $0.05 \leq P < 0.1$; see Table S1) and all traits except plate number in *EDA* heterozygotes. The latter trait was excluded because mapping was performed with substantially reduced power ($N = 209$).

2009; Bergendal et al. 2011)—the same disorder also observed for disruptions of the *EDA* pathway (Kere et al. 1996; Bayes et al. 1998; Monreal et al. 1999; Headon et al. 2001; Chassaing et al. 2006). The SNP interval around 6.2 Mb on the same chromosome, in turn, proved close to *PHOSPHO1*, a gene involved in skeletal tissue mineralization (Houston et al. 1999, 2004; Roberts et al. 2007). Candidate genes involved in bone formation also emerged at the suggestive plate height QTLs on chromosomes 5 (*PLEKHM1*; Van Wesenbeeck et al. 2007) and 17 (*ALPL*; Weiss et al. 1988; Henthorn et al. 1992; Table S1).

QTL EFFECT SIZES

Mapping lateral plate morph and snout length identified QTLs of very large effect (acknowledging that the effect size of the plate morph QTL is possibly confounded by the presence of multiple tightly linked loci, see above). The majority of our detected QTLs, however, had a relatively minor effect, a typical result in QTL mapping studies (Mackay et al. 2009). Moreover, for some traits (e.g., gill raker length, head length), inspecting the genome-wide LOD profile (Fig. S1) suggested the presence of additional loci with even smaller effect that were missed in our experiment due to insufficient power—a well-known issue in QTL mapping (Lande and Thompson 1990; Beavis 1994; Xu 2003; Rockman 2011). To explore this issue further, we compiled the effect sizes of all the QTLs (including marginally significant ones) detected in our study across all traits. This revealed that despite relatively large F_2 sample size, we lacked the power to identify QTLs with an effect size below 3–4 PVE (Fig. 5). (We are aware of the additional complication that the effect sizes of our detected *minor* QTLs are likely to be biased upward; Beavis 1994; Göring et al. 2001; Xu 2003.) We thus argue that although QTL mapping provides interesting insights into the genetic architecture of phenotypic divergence among stickleback populations, our understanding of adaptive variation in many traits will continue to benefit from quantitative genetic investigation.

ALLELE FREQUENCY SHIFTS IN THE SOURCE POPULATIONS

The availability of marker data from the natural populations underlying our cross made it possible to assess if associations between trait values and QTL-linked SNP alleles were replicated at the population level. Such an association is expected under two conditions. First, SNP alleles must tag QTL alleles reliably at the population level (as opposed to merely in the grandparents used for the cross). Second, allele frequency shifts at the focal QTL need to make some contribution to the trait divergence between the natural populations. Moreover, a shift at an SNP tightly linked to a QTL is expected to exceed the level of baseline population divergence attributable to drift if the QTL has been influenced by divergent selection between the populations.

All these conditions are indeed met by the *EDA* locus: phylogenetic analysis revealed that alleles at SNPs within *EDA* are tightly linked to their corresponding causative variants (which remain unknown) in the two study populations, and that adaptive population divergence in plate morph frequency is paralleled by frequency shifts at these SNPs (Berner et al. 2010a; see also Colosimo et al. 2005). We thus predicted very strong population-level shifts at our *EDA* marker, which were indeed observed. Similar analyses could be performed only in a small subset of the other QTL-linked SNPs, because some relevant markers were missing at the population level due to a different sequencing protocol, and because plate height could not be quantified in the (low-plated) stream population. Nevertheless, all three additional SNPs that were examined (associated with gill raker length, head length, and vertebral number) showed clear enrichment for the expected allele within each source population. Moreover, in three of the four total cases (including *EDA*), the observed allele frequency shifts were stronger than baseline divergence. We thus conclude that the phenotypic divergence between our study populations is probably attributable at least partly to allele frequency shifts at the QTL discovered in the cross, and that some of these shifts have been driven by divergent selection.

Conclusions

We subjected skeletal traits in European lake and stream stickleback to QTL mapping. Although this revealed a few large-effect QTLs, the majority of the loci detected across all traits exhibited a modest to small effect size. At least for some traits, QTL mapping seems to permit a relatively incomplete characterization of genetic architecture. Nevertheless, the close neighborhood around the QTLs that were discovered often contained genes involved in bone formation, which thus emerge as strong candidate drivers of skeletal evolution. Manipulative functional experiments are now needed to confirm the causative role of these genes, and

comparisons across numerous phenotypically well-characterized stickleback populations should investigate how consistently these genes are involved in diversification. Excitingly, a region containing two novel candidate genes for lateral plate morph evolution in our study coincided with an outlier region identified in a divergence scan using geographically independent stickleback populations divergent in lateral plating (Hohenlohe et al. 2010). This illustrates how understanding adaptation can benefit from the combination of phenotype-based and purely molecular genome scans. Finally, we attempted to move beyond mere QTL identification within a cross by screening for QTL-linked SNP allele frequency shifts in the natural source populations. These analyses indicated that at least some of our identified QTLs may indeed contribute to population divergence, and suggested that allele frequency shifts have been driven by divergent selection. A deeper understanding of the nature of this selection, however, will require extensive ecological investigation.

ACKNOWLEDGMENTS

We gratefully acknowledge the following contributions: W. Cresko and C. Peichel provided input on the cross and marker generation design. F. Hofmann (SFFN—Inspection de la pêche VD) and R. Kistler (fisheries authorities of the canton Thurgau) provided sampling permits for the cross populations, and A.-C. Grandchamp aided field sampling. B. Egger, H. Gante, A. Indermaur, A. Theis, and P. Vonlanthen aided fish husbandry. X-ray scanning was supported by the group of M. Kneissel, Global Head Musculoskeletal Disease Area, Novartis Institutes for BioMedical Research, Basel, Switzerland. L. Bänziger, A. Frey, B. Kueng, and S. Moser carried out preliminary analyses. B. Aeschbach and N. Boileau facilitated wet laboratory work. Illumina sequencing was done by I. Nissen and C. Beisel at the Quantitative Genomics Facility, D-BSSE, ETH Zürich. Novocraft (<http://www.novocraft.com>) shared their sequence aligner. J. Merilä, C. Peichel, and an anonymous reviewer provided valuable comments on the manuscript. This study was funded by the Swiss National Science Foundation (grant 31003A.146208/1 and Ambizione fellowship PZ00P3.126391/1 to DB; Sinergia grant CRSII3.136293 to WS), the Freiwillige Akademische Gesellschaft (FAG) Basel (DB), and the Research Fund of the University of Basel (DB, WS). DB designed the study, produced the cross, measured phenotypes, generated the genotype data, performed all mapping, and wrote the paper, with input from all co-authors. DM contributed to fish husbandry. DM and MR did the molecular wet lab work and measured phenotypes. MR contributed to study design and did the candidate gene search. HB generated all X-ray scans. WS contributed to study design and provided resources and infrastructure.

DATA ARCHIVING

The doi for our data is 10.5061/dryad.b2534.

LITERATURE CITED

- Aguirre, W. E. 2009. Microgeographical diversification of threespine stickleback: body shape-habitat correlations in a small, ecologically diverse Alaskan drainage. *Biol. J. Linn. Soc.* 98:139–151.
- Aguirre, W. E., and O. Akinpelu. 2010. Sexual dimorphism of head morphology in threespine stickleback (*Gasterosteus aculeatus*). *J. Fish Biol.* 77:802–821.

- Akey, J. M., G. Zhang, K. Zhang, L. Jin, and M. D. Shriver. 2002. Interrogating a high-density SNP map for signatures of natural selection. *Genome Res.* 12:1805–1814.
- Albert, A. Y. K., S. Sawaya, T. H. Vines, A. K. Knecht, C. T. Miller, B. R. Summers, S. Balabhadra, D. M. Kingsley, and D. Schluter. 2008. The genetics of adaptive shape shift in stickleback: pleiotropy and effect size. *Evolution* 62:76–85.
- Alho, J. S., T. Leinonen, and J. Merilä. 2011. Inheritance of vertebral number in the three-spined stickleback (*Gasterosteus aculeatus*). *PLoS One* 6:e19579.
- 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, E. U. Selker, W. A. Cresko, and E. A. Johnson. 2008. Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS One* 3:e3376.
- Barrett, R. D. H., S. M. Rogers, and D. Schluter. 2009. Environment specific pleiotropy facilitates divergence at the *ectodysplasin* locus in threespine stickleback. *Evolution* 63:2831–2837.
- Barton, N. H., and P. D. Keightley. 2002. Understanding quantitative genetic variation. *Nat. Rev. Genet.* 3:11–21.
- Bayes, M., A. J. Hartung, S. Ezer, J. Pispá, I. Thesleff, A. K. Srivastava, and J. Kere. 1998. The anhidrotic ectodermal dysplasia gene (EDA) undergoes alternative splicing and encodes ectodysplasin-A with deletion mutations in collagenous repeats. *Hum. Mol. Genet.* 7:1661–1669.
- Beavis, W. D. 1994. The power and deceit of QTL experiments: lessons from comparative QTL studies. Pp. 250–266 in *Proceedings of the 49th annual corn and sorghum industry research conference*. American Seed Trade Association, Washington, DC.
- Bell, M. A., and S. A. Foster. 1994. *The evolutionary biology of the threespine stickleback*. Oxford Univ. Press, Oxford, U.K.
- Bentzen, P., and J. D. McPhail. 1984. Ecology and evolution of sympatric sticklebacks (*Gasterosteus*): specialization for alternative trophic niches in the Enos Lake species pair. *Can. J. Zool.* 62:2280–2286.
- Bentzen, P., M. S. Ridgway, and J. D. McPhail. 1984. Ecology and evolution of sympatric sticklebacks (*Gasterosteus*)—spatial segregation and seasonal habitat shifts in the Enos Lake species pair. *Can. J. Zool.* 62:2436–2439.
- Bergendal, B., J. Klar, C. Stecksén-Blicks, J. Norderyd, and N. Dahl. 2011. Isolated oligodontia associated with mutations in EDARADD, AXIN2, MSX1, and PAX9 genes. *Am. J. Med. Genet. Part A* 155A:1616–1622.
- Bergstrom, C. A. 2002. Fast-start swimming performance and reduction in lateral plate number in threespine stickleback. *Can. J. Zool.* 80:207–213.
- Berner, D. 2011. Size correction in biology: how reliable are approaches based on (common) principal component analysis? *Oecologia* 166:961–971.
- Berner, D., D. C. Adams, A.-C. Grandchamp, and A. P. Hendry. 2008. Natural selection drives patterns of lake-stream divergence in stickleback foraging morphology. *J. Evol. Biol.* 21:1653–1665.
- Berner, D., A.-C. Grandchamp, and A. P. Hendry. 2009. Variable progress toward ecological speciation in parapatry: stickleback across eight lake-stream transitions. *Evolution* 63:1740–1753.
- Berner, D., M. Roesti, A. P. Hendry, and W. Salzburger. 2010a. Constraints on speciation suggested by comparing lake-stream stickleback divergence across two continents. *Mol. Ecol.* 19:4963–4978.
- Berner, D., W. E. Stutz, and D. I. Bolnick. 2010b. Foraging trait (co)variances in stickleback evolve deterministically and do not predict trajectories of adaptive diversification. *Evolution* 64:2265–2277.
- Berner, D., R. Kaeuffer, A.-C. Grandchamp, J. A. M. Raeymaekers, K. Räsänen, and A. P. Hendry. 2011. Quantitative genetic inheritance of morphological divergence in a lake-stream stickleback ecotype pair: implications for reproductive isolation. *J. Evol. Biol.* 24:1975–1983.
- Boden, S. D., Y. S. Liu, G. A. Hair, J. A. Helms, D. Hu, M. Racine, M. S. Nanes, and L. Titus. 1998. LMP-1, a LIM-domain protein, mediates BMP-6 effects on bone formation. *Endocrinology* 139:5125–5134.
- Bolnick, D. I. 2004. Can intraspecific competition drive disruptive selection? An experimental test in natural populations of sticklebacks. *Evolution* 58:608–618.
- Bolnick, D. I., L. K. Snowberg, C. Pattenia, W. E. Stutz, T. Ingram, and O. L. Lau. 2009. Phenotype-dependent native habitat preference facilitates divergence between parapatric lake and stream stickleback. *Evolution* 63:2004–2016.
- Broman, K. W., and S. Sen. 2009. *A guide to QTL mapping with R/qtl*. Springer, New York.
- Caldecutt, W. J., and D. C. Adams. 1998. Morphometrics of trophic osteology in the threespine stickleback, *Gasterosteus aculeatus*. *Copeia* 1998:827–838.
- Callahan, N., A. Modesto, R. Meira, F. Seymen, A. Patir, and A. R. Vieira. 2009. Axis inhibition protein 2 (AXIN2) polymorphisms and tooth agenesis. *Arch. Oral Biol.* 54:45–49.
- Chassaing, N., S. Bourthoumieu, M. Cosse, P. Calvas, and M. C. Vincent. 2006. Mutations in EDAR account for one-quarter of non-ED1-related hypohidrotic ectodermal dysplasia. *Hum. Mutat.* 27:255–259.
- Colosimo, P. F., C. L. Peichel, K. Nereng, B. K. Blackman, M. D. Shapiro, D. Schluter, and D. M. Kingsley. 2004. The genetic architecture of parallel armor plate reduction in threespine sticklebacks. *PLoS Biol.* 2:635–641.
- Colosimo, P. F., K. E. Hosemann, S. Balabhadra, G. Villareal Jr., M. Dickson, J. Grimwood, J. Schmutz, R. M. Myers, D. Schluter, and D. M. Kingsley. 2005. Widespread parallel evolution in sticklebacks by repeated fixation of ectodysplasin alleles. *Science* 307:1928–1933.
- Cresko, W. A., A. Amores, C. Wilson, J. Murphy, M. Currey, P. Phillips, M. A. Bell, C. B. Kimmel, and J. H. Postlethwait. 2004. Parallel genetic basis for repeated evolution of armor loss in Alaskan threespine stickleback populations. *Proc. Natl. Acad. Sci. USA* 101:6050–6055.
- Day, T., J. Pritchard, and D. Schluter. 1994. Ecology and genetics of phenotypic plasticity: a comparison of two sticklebacks. *Evolution* 48:1723–1734.
- Deagle, B. E., F. C. Jones, Y. F. Chan, D. M. Absher, D. M. Kingsley, and T. E. Reimchen. 2012. Population genomics of parallel phenotypic evolution in stickleback across stream-lake ecological transitions. *Proc. R. Soc. Lond. B* 279:1277–1286.
- Gerking, S. D. 1994. *Feeding ecology of fish*. Academic, San Diego, CA.
- 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.
- Göring, H. H. H., J. D. Terwilliger, and J. Blangero. 2001. Large upward bias in estimation of locus-specific effects from genomewide scans. *Am. J. Hum. Genet.* 69:1357–1369.
- Gross, H. P., and J. M. Anderson. 1984. Geographic variation in the gillrakers and diet of European sticklebacks, *Gasterosteus aculeatus*. *Copeia* 1984:87–97.
- Hagen, D. W., and L. G. Gilbertson. 1972. Geographic variation and environmental selection in *Gasterosteus aculeatus* L. in the Pacific Northwest, America. *Evolution* 26:32–51.
- Harris, M. P., N. Rohner, H. Schwarz, S. Perathoner, P. Konstantinidis, and C. Nüsslein-Volhard. 2008. Zebrafish *eda* and *edar* mutants reveal conserved and ancestral roles of ectodysplasin signaling in vertebrates. *PLoS Genet.* 4:e1000206.
- Headon, D. J., S. A. Emmal, B. M. Ferguson, A. S. Tucker, M. J. Justice, P. T. Sharpe, J. Zonana, and P. A. Overbeek. 2001. Gene defect in ectodermal

- dysplasia implicates a death domain adapter in development. *Nature* 414:913–916.
- Hendry, A. P., and E. B. Taylor. 2004. How much of the variation in adaptive divergence can be explained by gene flow? An evaluation using lake-stream stickleback pairs. *Evolution* 58:2319–2331.
- Henthorn, P. S., M. Raducha, K. N. Fedde, M. A. Lafferty, and M. P. Whyte. 1992. Different missense mutations at the tissue-nonspecific alkaline-phosphatase gene locus in autosomal recessively inherited forms of mild and severe hypophosphatasia. *Proc. Natl. Acad. Sci. USA* 89:9924–9928.
- Heuts, M. J. 1947. Experimental studies on adaptive evolution in *Gasterosteus aculeatus* L. *Evolution* 1:89–102.
- Hoekstra, H. E., and J. A. Coyne. 2007. The locus of evolution: evo devo and the genetics of adaptation. *Evolution* 61:995–1016.
- 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.
- Houston, B., E. Seawright, D. Jefferies, E. Hoogland, D. Lester, C. Whitehead, and C. Farquharson. 1999. Identification and cloning of a novel phosphatase expressed at high levels in differentiating growth plate chondrocytes. *Biochim. Biophys. Acta* 1448:500–506.
- Houston, B., A. J. Stewart, and C. Farquharson. 2004. PHOSPHO1—a novel phosphatase specifically expressed at sites of mineralisation in bone and cartilage. *Bone* 34:629–637.
- Jones, F. C., M. G. Grabherr, Y. F. Chan, P. Russell, E. Mauceli, J. Johnson, R. Swofford, M. Pirun, M. C. Zody, S. White, et al. 2012. The genomic basis of adaptive evolution in threespine sticklebacks. *Nature* 484:55–61.
- Kaeuffer, R., C. Peichel, D. I. Bolnick, and A. P. Hendry. 2012. Convergence and non-convergence in ecological, phenotypic, and genetic divergence across replicate population pairs of lake and stream stickleback. *Evolution* 66:402–418.
- Kahilainen, K., and K. Ostbye. 2006. Morphological differentiation and resource polymorphism in three sympatric whitefish *Coregonus lavaretus* (L.) forms in a subarctic lake. *J. Fish Biol.* 68:63–79.
- Kalamajski, S., A. Aspberg, K. Lindblom, D. Heinegard, and A. Oldberg. 2009. Asporin competes with decorin for collagen binding, binds calcium and promotes osteoblast collagen mineralization. *Biochem. J.* 423:53–59.
- Kere, J., A. K. Srivastava, O. Montonen, J. Zonana, N. Thomas, B. Ferguson, F. Munoz, D. Morgan, A. Clarke, P. Baybayan, et al. 1996. X-linked anhidrotic (hypohidrotic) ectodermal dysplasia is caused by mutation in a novel transmembrane protein. *Nat. Genet.* 13:409–416.
- Kimura, T., M. Shinya, and K. Naruse. 2012. Genetic analysis of vertebral regionalization and number in Medaka (*Oryzias latipes*) inbred lines. *G3* 2:1317–1323.
- Kirsch, T., G. Harrison, E. E. Golub, and H. D. Nah. 2000. The roles of annexins and types II and X collagen in matrix vesicle-mediated mineralization of growth plate cartilage. *J. Biol. Chem.* 275:35577–35583.
- Kitano, J., S. Mori, and C. L. Peichel. 2007. Sexual dimorphism in the external morphology of the threespine stickleback (*Gasterosteus aculeatus*). *Copeia* 2007:336–349.
- Kizawa, H., I. Kou, A. Iida, A. Sudo, Y. Miyamoto, A. Fukuda, A. Mabuchi, A. Kotani, A. Kawakami, S. Yamamoto, et al. 2005. An aspartic acid repeat polymorphism in asporin inhibits chondrogenesis and increases susceptibility to osteoarthritis. *Nat. Genet.* 37:138–144.
- Koyama, K., K. Nakazato, S. K. Min, K. Gushiken, Y. Hatakeda, K. Seo, and K. Hiranuma. 2012. COL11A1 gene is associated with limbus vertebra in gymnasts. *Int. J. Sports Med.* 33:586–590.
- Lammi, L., S. Arte, M. Somer, H. Jarvinen, P. Lahermo, I. Thesleff, S. Pirinen, and P. Nieminen. 2004. Mutations in AXIN2 cause familial tooth agenesis and predispose to colorectal cancer. *Am. J. Hum. Genet.* 74:1043–1050.
- Lande, R., and R. Thompson. 1990. Efficiency of marker-assisted selection in the improvement of quantitative traits. *Genetics* 124:743–756.
- Lavin, P. A., and J. D. McPhail. 1993. Parapatric lake and stream sticklebacks on northern Vancouver Island: disjunct distribution or parallel evolution? *Can. J. Zool.* 71:11–17.
- Lawniczak, M. K. N., S. J. Emrich, A. K. Holloway, A. P. Regier, M. Olson, B. White, S. Redmond, L. Fulton, E. Appelbaum, J. Godfrey, et al. 2010. Widespread divergence between incipient *Anopheles gambiae* species revealed by whole genome sequences. *Science* 330:512–514.
- Leinonen, T., G. Herczeg, J. M. Cano, and J. Merilä. 2011a. Predation-imposed selection on threespine stickleback (*Gasterosteus aculeatus*) morphology: a test of the refuge use hypothesis. *Evolution* 65:2916–2926.
- Leinonen, T., J. M. Cano, and J. Merilä. 2011b. Genetics of body shape and armour variation in threespine sticklebacks. *J. Evol. Biol.* 24:206–218.
- Leinonen, T., R. J. S. McCairns, G. Herczeg, and J. Merilä. 2012. Multiple evolutionary pathways to decreased lateral plate coverage in freshwater threespine sticklebacks. *Evolution* 66:3866–3875.
- Lessells, C. M., and P. T. Boag. 1987. Unrepeatable repeatabilities—a common mistake. *Auk* 104:116–121.
- Li, Y., D. A. Lacerda, M. L. Warman, D. R. Beier, H. Yoshioka, Y. Ninomiya, J. T. Oxford, N. P. Morris, K. Andrikopoulos, F. Ramirez, et al. 1995. A fibrillar collagen gene, COL11A1, is essential for skeletal morphogenesis. *Cell* 80:423–430.
- Liu, Y. H., G. A. Hair, S. D. Boden, M. Viggswarapu, and L. Titus. 2002. Overexpressed LIM mineralization proteins do not require LIM domains to induce bone. *J. Bone Miner. Res.* 17:406–414.
- Lucek, K., A. Sivasundar, and O. Seehausen. 2012. Evidence of adaptive evolutionary divergence during biological invasion. *PLoS One* 7:e49377.
- Lucek, K., A. Sivasundar, D. Roy, and O. Seehausen. 2013. Repeated and predictable patterns of ecotypic differentiation during a biological invasion: lake-stream divergence in parapatric Swiss stickleback. *J. Evol. Biol.* 26:2691–2709.
- Mackay, T. F. C., E. A. Stone, and J. F. Ayroles. 2009. The genetics of quantitative traits: challenges and prospects. *Nat. Rev. Genet.* 10:565–577.
- Madisen, L., M. Neubauer, G. Plowman, D. Rosen, P. Segarini, J. Dasch, A. Thompson, J. Ziman, H. Bentz, and A. F. Purchio. 1990. Molecular cloning of a novel bone-forming compound—osteoinductive factor. *DNA Cell Biol.* 9:303–309.
- Manly, B. F. J. 2007. Randomization, bootstrap and Monte Carlo methods in biology. Chapman & Hall, Boca Raton, FL.
- Mateus, C. S., M. Stange, D. Berner, M. Roesti, B. R. Quintella, M. J. Alves, P. R. Almeida, and W. Salzburger. 2013. Strong genome-wide divergence between sympatric European river and brook lampreys. *Curr. Biol.* 23:R649–R650.
- Matthews, B., K. B. Marchinko, D. I. Bolnick, and A. Mazumder. 2010. Specialization of trophic position and habitat use by sticklebacks in an adaptive radiation. *Ecology* 91:1025–1034.
- McDowall, R. M. 2008. Jordan's and other ecogeographical rules, and the vertebral number in fishes. *J. Biogeogr.* 35:501–508.
- McPhail, J. D. 1984. Ecology and evolution of sympatric sticklebacks (*Gasterosteus*): morphological and genetic evidence for a species pair in Enos Lake, British Columbia. *Can. J. Zool.* 62:1402–1408.
- Mikawa, S., T. Morozumi, S.-I. Shimanuki, T. Hayashi, H. Uenishi, M. Domukai, N. Okumura, and T. Awata. 2007. Fine mapping of a swine quantitative trait locus for number of vertebrae and analysis of an orphan nuclear receptor, germ cell nuclear factor (NR6A1). *Genome Res.* 17:586–593.

- Mikawa, S., S. Sato, M. Nii, T. Morozumi, G. Yoshioka, N. Imaeda, T. Yamaguchi, T. Hayashi, and T. Awata. 2011. Identification of a second gene associated with variation in vertebral number in domestic pigs. *BMC Genet.* 12:5.
- Miller, C. T., D. Yelon, D. Y. R. Stainier, and C. B. Kimmel. 2003. Two *endothelin 1* effectors, *hand2* and *bapx1*, pattern ventral pharyngeal cartilage and the jaw joint. *Development* 130:1353–1365.
- Mitchell-Olds, T., J. H. Willis, and D. B. Goldstein. 2007. Which evolutionary processes influence natural genetic variation for phenotypic traits? *Nat. Rev. Genet.* 8:845–856.
- Monreal, A. W., B. M. Ferguson, D. J. Headon, S. L. Street, P. A. Overbeek, and J. Zonana. 1999. Mutations in the human homologue of mouse *dl* cause autosomal recessive and dominant hypohidrotic ectodermal dysplasia. *Nat. Genet.* 22:366–369.
- Moodie, G. E. E., and T. E. Reimchen. 1976. Phenetic variation and habitat differences in *Gasterosteus* populations of Queen Charlotte Islands. *Syst. Zool.* 25:49–61.
- Moser, D., M. Roesti, and D. Berner. 2012. Repeated lake-stream divergence in stickleback life history within a Central European lake basin. *PLoS One* 7:e50620.
- Mostowska, A., B. Biedziak, and P. P. Jagodzinski. 2006. Axis inhibition protein 2 (*AXIN2*) polymorphisms may be a risk factor for selective tooth agenesis. *J. Hum. Genet.* 51:262–266.
- Myhre, F., and T. Klepaker. 2009. Body armour and lateral-plate reduction in freshwater three-spined stickleback *Gasterosteus aculeatus*: adaptations to a different buoyancy regime? *J. Fish Biol.* 75:2062–2074.
- Nadeau, N. J., A. Whibley, R. T. Jones, J. W. Davey, K. K. Dasmahapatra, S. W. Baxter, M. A. Quail, M. Joron, R. H. French-Constant, M. L. Blaxter, et al. 2012. Genomic islands of divergence in hybridizing *Heliconius* butterflies identified by large-scale targeted sequencing. *Philos. Trans. R. Soc. Lond. B* 367:343–353.
- Nei, M., and F. Tajima. 1981. DNA polymorphism detectable by restriction endonucleases. *Genetics* 97:145–163.
- Nichols, K. M., P. A. Wheeler, and G. H. Thorgaard. 2004. Quantitative trait loci analyses for meristic traits in *Oncorhynchus mykiss*. *Environ. Biol. Fish.* 69:317–331.
- Nielsen, R. 2005. Molecular signatures of natural selection. *Annu. Rev. Genet.* 39:197–218.
- Oleksyk, T. K., M. W. Smith, and S. J. O'Brien. 2010. Genome-wide scans for footprints of natural selection. *Philos. Trans. R. Soc. Lond. B* 365:185–205.
- Orr, H. A. 1998. The population genetics of adaptation: the distribution of factors fixed during adaptive evolution. *Evolution* 52:935–949.
- Pfaender, J., F. W. Miesen, R. K. Hadiaty, and F. Herder. 2011. Adaptive speciation and sexual dimorphism contribute to diversity in form and function in the adaptive radiation of Lake Matano's sympatric roundfin sailfin silversides. *J. Evol. Biol.* 24:2329–2345.
- Phillips, P. C. 2005. Testing hypotheses regarding the genetics of adaptation. *Genetica* 123:15–24.
- R Development Core Team 2013. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Austria, Vienna.
- Ravinet, M., P. A. Prodoehl, and C. Harrod. 2013. Parallel and nonparallel ecological, morphological and genetic divergence in lake-stream stickleback from a single catchment. *J. Evol. Biol.* 26:186–204.
- Reimchen, T. E. 1994. Predators and morphological evolution in threespine stickleback. Pp. 240–273 in M. A. Bell and S. A. Foster, eds. *The evolutionary biology of the threespine stickleback*. Oxford Univ. Press, Oxford, U.K.
- Reimchen, T. E., E. M. Stinson, and J. S. Nelson. 1985. Multivariate differentiation of parapatric and allopatric populations of threespine stickleback in the Sangan River watershed, Queen Charlotte Islands. *Can. J. Zool.* 63:2944–2951.
- Reist, J. D. 1985. An empirical evaluation of several univariate methods that adjust for size variation in morphometric data. *Can. J. Zool.* 63:1429–1439.
- Ren, D. R., J. Ren, G. F. Ruan, Y. M. Guo, L. H. Wu, G. C. Yang, L. H. Zhou, L. Li, Z. Y. Zhang, and L. S. Huang. 2012. Mapping and fine mapping of quantitative trait loci for the number of vertebrae in a White Duroc x Chinese Erhualian intercross resource population. *Anim. Genet.* 43:545–551.
- Roberts, S., S. Narisawa, D. Harmey, J. L. Millan, and C. Farquharson. 2007. Functional involvement of PHOSPHO1 in matrix vesicle-mediated skeletal mineralization. *J. Bone Miner. Res.* 22:617–627.
- Rockman, M. V. 2011. The QTN program and the alleles that matter for evolution: all that's gold does not glitter. *Evolution* 66:1–17.
- Roesti, M., A. P. Hendry, W. Salzburger, and D. Berner. 2012a. Genome divergence during evolutionary diversification as revealed in replicate lake-stream stickleback population pairs. *Mol. Ecol.* 21:2852–2862.
- Roesti, M., W. Salzburger, and D. Berner. 2012b. Uninformative polymorphisms bias genome scans for signatures of selection. *BMC Evol. Biol.* 12:94.
- Roesti, M., D. Moser, and D. Berner. 2013. Recombination in the threespine stickleback genome—patterns and consequences. *Mol. Ecol.* 22:3014–3027.
- 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. In press*.
- Rohlf, F. J. 2001. tpsDig, tpsRelw, tpsRegr. Distributed by the author. Department of Ecology and Evolution, State University of New York, Stony Brook, NY.
- Sanderson, S. L., A. Y. Cheer, J. S. Goodrich, J. D. Graziano, and W. T. Callan. 2001. Crossflow filtration in suspension-feeding fishes. *Nature* 412:439–441.
- Schluter, D., and J. D. McPhail. 1992. Ecological character displacement and speciation in sticklebacks. *Am. Nat.* 140:85–108.
- Sharpe, D. M. T., K. Räsänen, D. Berner, and A. P. Hendry. 2008. 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.
- Song, Y.-Q., K. M. C. Cheung, D. W. H. Ho, S. C. S. Poon, K. Chiba, Y. Kawaguchi, Y. Hirose, M. Alini, S. Grad, A. F. Y. Yee, et al. 2008. Association of the asporin D14 allele with lumbar-disc degeneration in Asians. *Am. J. Hum. Genet.* 82:744–747.
- Stern, D. L., and V. Orgogozo. 2008. The loci of evolution: how predictable is genetic evolution? *Evolution* 62:2155–2177.
- Stinchcombe, J. R., and H. E. Hoekstra. 2008. Combining population genomics and quantitative genetics: finding the genes underlying ecologically important traits. *Heredity* 100:158–170.
- Stölting, K. N., R. Nipper, D. Lindtke, C. Caseys, S. Waeber, S. Castiglione, and C. Lexer. 2013. Genomic scan for single nucleotide polymorphisms reveals patterns of divergence and gene flow between ecologically divergent species. *Mol. Ecol.* 22:842–855.
- Storz, J. F. 2005. Using genome scans of DNA polymorphism to infer adaptive population divergence. *Mol. Ecol.* 14:671–688.
- Storz, J. F., and C. W. Wheat. 2010. Integrating evolutionary and functional approaches to infer adaptation at specific loci. *Evolution* 64:2489–2509.
- Swain, D. P. 1992. The functional basis of natural selection for vertebral traits of larvae in the stickleback *Gasterosteus aculeatus*. *Evolution* 46:987–997.
- The UniProt Consortium. 2013. Update on activities at the Universal Protein Resource (UniProt) in 2013. *Nucleic Acids Res.* 41:D43–D47.

- Thompson, C. E., E. B. Taylor, and J. D. McPhail. 1997. Parallel evolution of lake-stream pairs of threespine sticklebacks (*Gasterosteus*) inferred from mitochondrial DNA variation. *Evolution* 51:1955–1965.
- Thouverey, C., A. Strzelecka-Kiliszek, M. Balcerzak, R. Buchet, and S. Pikula. 2009. Matrix vesicles originate from apical membrane microvilli of mineralizing osteoblast-like saos-2 cells. *J. Cell. Biochem.* 106:127–138.
- Tompson, S. W., C. A. Bacino, N. P. Safina, M. B. Bober, V. K. Proud, T. Funari, M. F. Wangler, L. Nevarez, L. Ala-Kokko, W. R. Wilcox, et al. 2010. Fibrochondrogenesis results from mutations in the COL11A1 type XI collagen gene. *Am. J. Hum. Genet.* 87:708–712.
- Van Wesenbeeck, L., P. R. Odgren, F. P. Coxon, A. Frattini, P. Moens, B. Perdu, C. A. MacKay, E. Van Hul, J.-P. Timmermans, F. Vanhoenacker, et al. 2007. Involvement of PLEKHM1 in osteoclastic vesicular transport and osteopetrosis in incisors absent rats and humans. *J. Clin. Invest.* 117:919–930.
- Voight, B. F., S. Kudravalli, X. Wen, and J. K. Pritchard. 2006. A map of recent positive selection in the human genome. *PLoS Biol.* 4:e72.
- Wake, D. B., M. H. Wake, and C. D. Specht. 2011. Homoplasy: from detecting pattern to determining process and mechanism of evolution. *Science* 331:1032–1035.
- Wang, W., and T. Kirsch. 2002. Retinoic acid stimulates annexin-mediated growth plate chondrocyte mineralization. *J. Cell Biol.* 157:1061–1069.
- Ward, A. B., and E. L. Brainerd. 2007. Evolution of axial patterning in elongate fishes. *Biol. J. Linn. Soc.* 90:97–116.
- Weiss, M. J., D. E. C. Cole, K. Ray, M. P. Whyte, M. A. Lafferty, R. A. Mulivor, and H. Harris. 1988. A missense mutation in the human liver/bone/kidney alkaline-phosphatase gene causing a lethal form of hypophosphatasia. *Proc. Natl. Acad. Sci. USA* 85:7666–7669.
- Wray, G. A. 2007. The evolutionary significance of cis-regulatory mutations. *Nat. Rev. Genet.* 8:206–216.
- Wund, M. A., J. A. Baker, B. Clancy, J. L. Golub, and S. A. Foster. 2008. A test of the “Flexible stem” model of evolution: ancestral plasticity, genetic accommodation, and morphological divergence in the threespine stickleback radiation. *Am. Nat.* 172:449–462.
- Xu, S. Z. 2003. Theoretical basis of the Beavis effect. *Genetics* 165:2259–2268.
- Yeaman, S., and M. C. Whitlock. 2011. The genetic architecture of adaptation under migration-selection balance. *Evolution* 65:1897–1911.

Associate Editor: C. Sgro

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Figure S1. Genome-wide LOD profiles for all eight traits.

Figure S2. Body size (quantified as geometric morphometric centroid size, in millimeters) in relation to vertebral number.

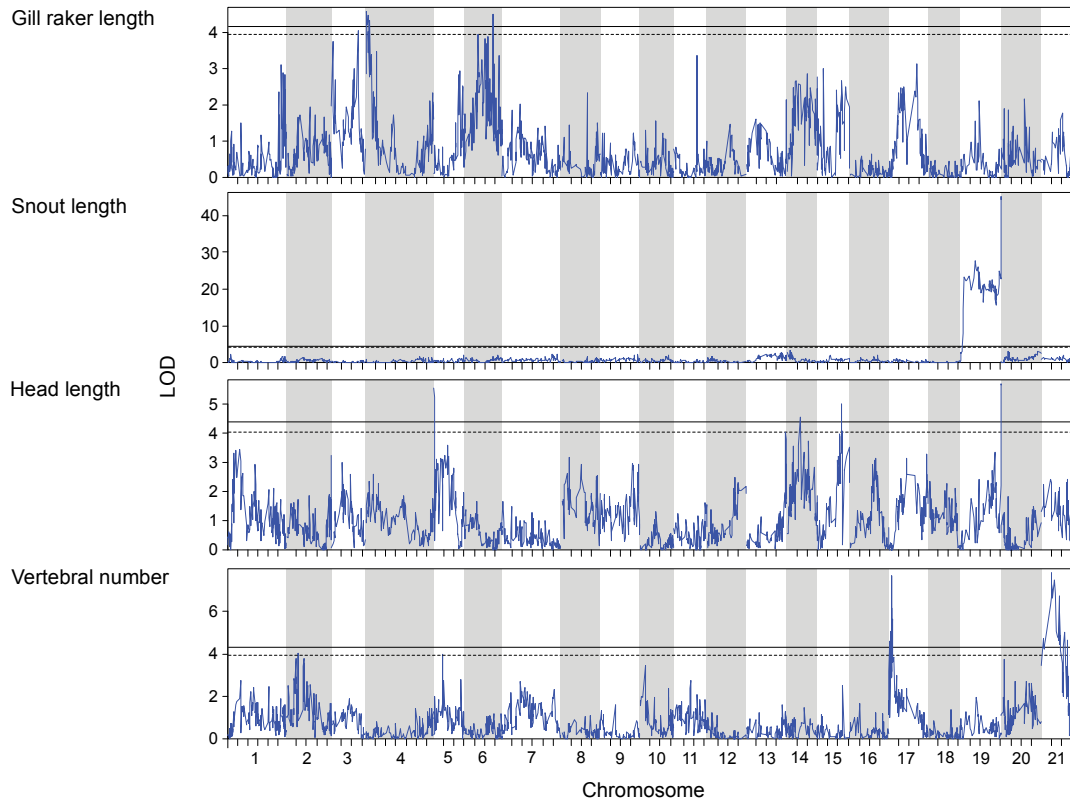
Table S1. Characterization of additional, suggestive ($0.05 \leq P < 0.1$) QTLs for skeletal divergence between lake and stream stickleback.

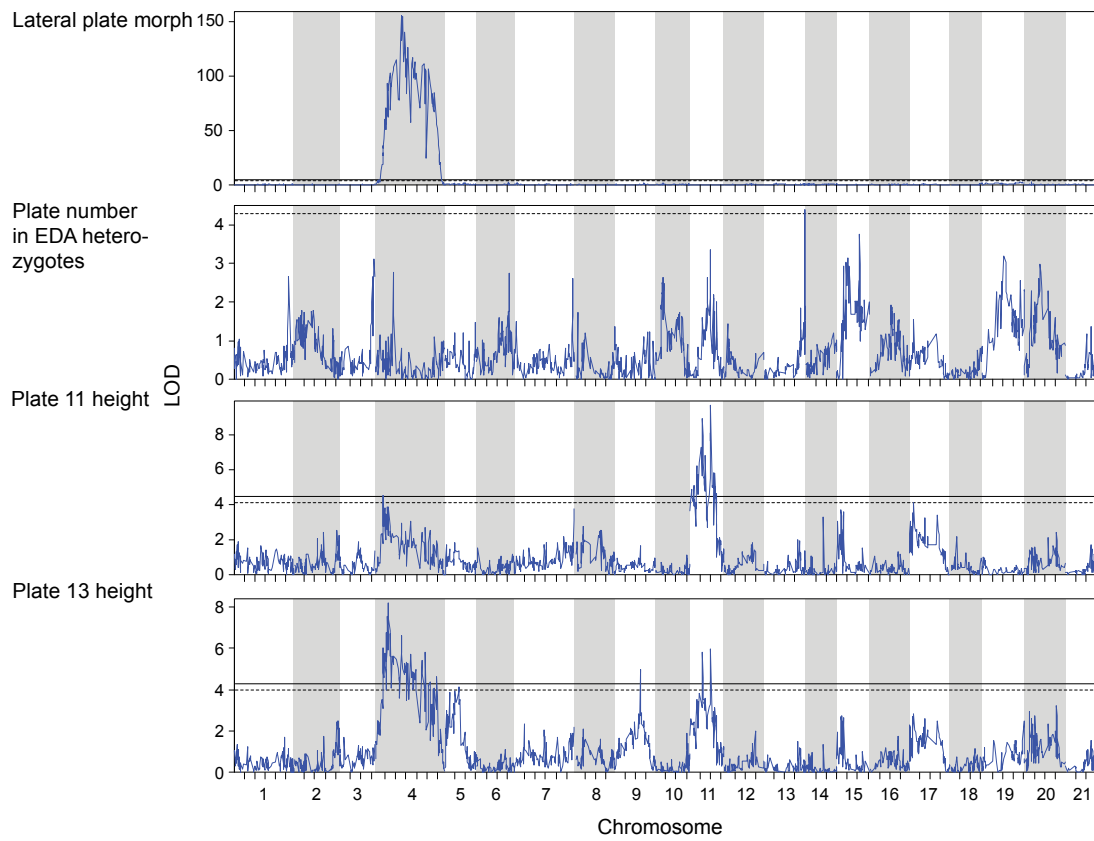
Chapter 5

Genetic architecture of skeletal evolution in European
lake and stream stickleback

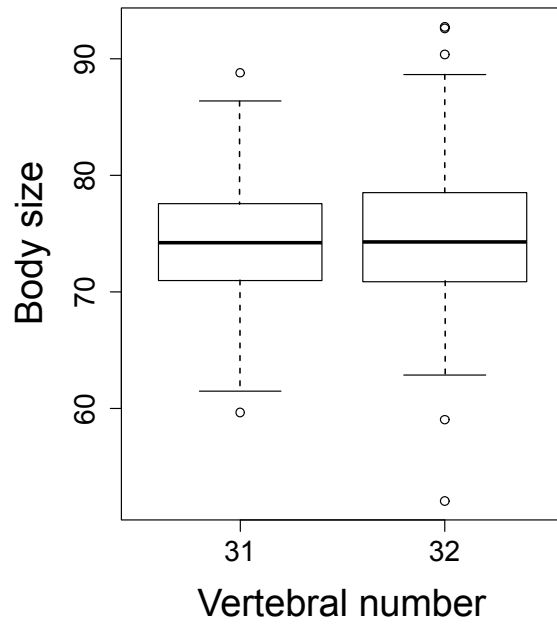
Supplementary material

Supporting Fig. S1. Genome-wide LOD profiles for all eight traits. The chromosomes are separated by white and gray background shading. Tick marks along the X-axis indicate 5 Mb intervals, drawn separately for each chromosome. Horizontal lines represent genome-wide 0.05 (solid) and 0.1 (dashed) LOD significance thresholds based on 1000 random permutations. Note that the Y-axes are scaled differently among the traits.





Supporting Fig. S2. Body size (quantified as geometric morphometric centroid size, in mm) in relation to vertebral number. Shown are only individuals with 31 (N = 197) and 32 (N = 258) vertebrae. The graph type is a boxplot with the whiskers representing 1.5 times the interquartile range.



Supporting Table S1. Characterization of additional, suggestive ($0.05 \leq P < 0.1$) QTL for skeletal divergence between lake and stream stickleback. The presentation follows Table 1. Plate number was mapped using only individuals unambiguously heterozygous at the Ectodysplasin (*EDA*) locus.

Trait	Marker	Chromosome	Position (Mb)	LOD	P	PVE	HSE	Direction	Candidate gene
Gill raker length	chrIII_13014992	3	13,014,992	4.06	0.07	4.4	0.14 (0.52)	L*	
Head length	chrXIII_18688122	13	19,444,640	4.05	0.086	4.6	0.40 (0.60)	S*	
Vertebral number	chrII_5665753	2	5,665,753	4.05	0.076	3.6	0.31 (0.57)	L*	
	chrV_6319659	5	4,509,573	3.96	0.089	3.6	0.11 (0.21)	L*	
Plate number	chrXIII_18503873	13	19,628,889	4.4	0.091	9.7	2.97 (0.75)	S	
Plate 11 height	chrXVII_2292374	17	2,292,374	4.12	0.098	5.8	0.36 (0.53)	L	<i>ALPL</i> (31)
Plate 13 height	chrV_3764951	5	7,064,281	4.15	0.069	5.6	0.33 (0.54)	L	<i>PLEKHM1</i> (12)

Chapter 6

The genomics of ecological vicariance in threespine
stickleback fish

ARTICLE

Received 1 Apr 2015 | Accepted 29 Sep 2015 | Published 10 Nov 2015

DOI: 10.1038/ncomms9767

OPEN

The genomics of ecological vicariance in threespine stickleback fish

Marius Roesti¹, Benjamin Kueng¹, Dario Moser¹ & Daniel Berner¹

Populations occurring in similar habitats and displaying similar phenotypes are increasingly used to explore parallel evolution at the molecular level. This generally ignores the possibility that parallel evolution can be mimicked by the fragmentation of an ancestral population followed by genetic exchange with ecologically different populations. Here we demonstrate such an ecological vicariance scenario in multiple stream populations of threespine stickleback fish divergent from a single adjacent lake population. On the basis of demographic and population genomic analyses, we infer the initial spread of a stream-adapted ancestor followed by the emergence of a lake-adapted population, that selective sweeps have occurred mainly in the lake population, that adaptive lake-stream divergence is maintained in the face of gene flow from the lake into the streams, and that this divergence involves major inversion polymorphisms also important to marine-freshwater stickleback divergence. Overall, our study highlights the need for a robust understanding of the demographic and selective history in evolutionary investigations.

¹Zoological Institute, University of Basel, Vesalgasse 1, Basel 4051, Switzerland. Correspondence and requests for materials should be addressed to M.R. (email: marius.roesti@unibas.ch) or to D.B. (email: daniel.berner@unibas.ch).

Parallel (or convergent¹) phenotypic evolution—that is, the repeated independent emergence of a specific phenotype associated with a specific habitat, can provide important insights into the determinism of natural selection. The reason is that similar phenotypes are unlikely to evolve repeatedly in association with an environment by chance. An aspect of parallel evolution now made amenable to investigation through advances in molecular techniques is to what extent the repeated evolution of similar phenotypes involves the same genetic loci^{1–3}. A common analytical framework adopted to address this question is to compare multiple population pairs, each believed to represent an independent replicate of adaptive population divergence between two ecologically different habitats. The evolutionary independence of these population pairs is generally established by demonstrating that the genetic relatedness between the populations within pairs, as inferred from markers little influenced by selection (for simplicity hereafter called ‘neutral markers’), exceeds that seen among the pairs. If so, the population pairs are assumed to represent replicates of independent ecological divergence and are screened for genomic loci exhibiting signatures of divergent selection between the habitats (for example, high divergence relative to some genome-wide baseline). Finally, the resulting lists of such loci are compared to draw conclusions about the extent of parallel evolution at the genomic level (for example, refs 4–9; for closely related inferential approaches see refs 10–12).

A possibility rarely considered in such investigations is that the demographic and selective history of the study populations may complicate or preclude inferences about parallel evolution. Such a situation occurs when multiple patches of two ecologically different habitats are initially colonized by a single ancestor already adapted to one habitat type. Subsequently, local adaptation in the alternative habitat drives ecologically based reproductive isolation between the habitats, although some genetic exchange across habitat boundaries will continue in the absence of absolute geographic barriers. The outcome of such ‘ecological vicariance’¹³ with genetic exchange will mimic parallel evolution¹⁴. The reason is that gene flow between ecologically different populations in contact will cause genetic differentiation at neutral markers to be lower within than among population pairs—the pattern also expected under parallel divergence. Moreover, under both scenarios, loci under divergent selection will be relatively protected from exchange between the populations in contact and can therefore maintain stronger differentiation between the habitats than neutral loci^{12,15–17}. In situations involving ecological vicariance with gene flow, comparing multiple population pairs can permit the reliable identification of selected loci and thus confirm divergent selection, but inference about the genetic basis of independent parallel evolution will be inappropriate because divergence did not occur repeatedly.

Distinguishing parallel divergence from ecological vicariance scenarios is thus crucial when attempting to explore how deterministically selection acts at the genomic level during evolution. While this distinction is not possible based on phylogenetic relationships at neutral markers^{18,19}, it can be achieved by combining thorough analyses of molecular signatures around the loci under divergent selection with robust reconstructions of the populations’ demographic history^{14,20}. We here present such an investigation based on populations of threespine stickleback fish (*Gasterosteus aculeatus*) adapted to lake and stream habitats within the Lake Constance basin in Central Europe.

This stickleback system comprises a large and genetically well-mixed population residing in Lake Constance—with 571 km² the third largest lake in Central Europe—and multiple adjoining

stream-resident populations inhabiting the lake’s tributaries^{21–23}. The lake and stream habitats are ecologically different, as mirrored by the lifestyles of the stickleback populations: lake fish forage pelagically (that is, in the open water) on zooplankton, whereas the stream populations feed on benthic (substrate-dwelling) macroinvertebrates. This different resource use is paralleled by divergence in foraging morphology and life history^{21,23,24}. Lake and stream populations in the Lake Constance basin also differ predictably in their extent of lateral plating^{21,23}. Just like marine stickleback²⁵, pelagic Lake Constance fish exhibit a series of bony plates covering their entire flank, providing protection from vertebrate predators in the open water²⁶. By contrast, multiple stream populations show a reduction in the extent of lateral plating, the phenotype predominant in freshwater stickleback on a global scale.

Although the Lake Constance stickleback system has certainly formed postglacially (that is, within the last 12,000 years²⁷), its origin is not resolved. One view is that a human introduction during the nineteenth century initially led to the establishment of a large lake population, and that subsequently multiple stream populations diverged independently from the lake population^{21,23}. This scenario thus implies parallel divergence. An alternative is a more ancient natural colonization of the Lake Constance region by an already stream-adapted ancestral population from the Danube drainage²³ (now draining into the Black Sea, hence disconnected from the Lake Constance basin), providing the potential for an ecological vicariance scenario.

The first goal of our study is to combine multiple lines of molecular evidence, based on dense genome-wide single-nucleotide polymorphisms (SNPs) obtained through restriction site-associated (RAD) sequencing²⁸, to resolve the demographic and selective history of lake–stream divergence in the Lake Constance stickleback system. We demonstrate that adaptive divergence has occurred in the face of gene flow in an unexpected historical context, pointing to limitations in the standard interpretation of repeated phenotypic evolution. Based on these insights, we then dissect the molecular consequences of divergent selection in target regions, including the prime locus underlying divergence in lateral plating, and finally examine the role of chromosomal inversions in adaptive divergence.

Results and Discussion

Demography and population genomic analyses. Our investigation focuses on four stickleback populations, including the panmictic (Supplementary Fig. 1) Lake Constance population (hereafter simply ‘lake’) and three stream populations residing in tributaries (referred to as Bohlingen (BOH), Nideraach (NID) and Grasbeuren (GRA); see also refs 21,23) (Fig. 1a), each represented by 22–25 individuals. To reconstruct the demographic history of these populations, we parameterized a divergence with gene flow model by using coalescent simulations based on the populations’ joint allele frequency spectra²⁹ derived from 14.8 million nucleotide positions on 166,711 RAD loci across the 460-Mb stickleback genome. This analysis indicated that the study populations—exhibiting relatively small estimated effective population sizes (extremely small in the lake, largest in GRA)—split from an at least 20 times larger ancestral population a few thousand generations (and years, since the typical life span of stickleback in this system is 1–2 years^{23,24}) ago (Fig. 1b). Qualitatively similar estimates were obtained with an alternative model including only two stream populations (Supplementary Fig. 2). Also, long-term rates of lake–stream gene flow differed approximately tenfold, being highest between the lake and the BOH population, and lowest between the lake and the GRA population.

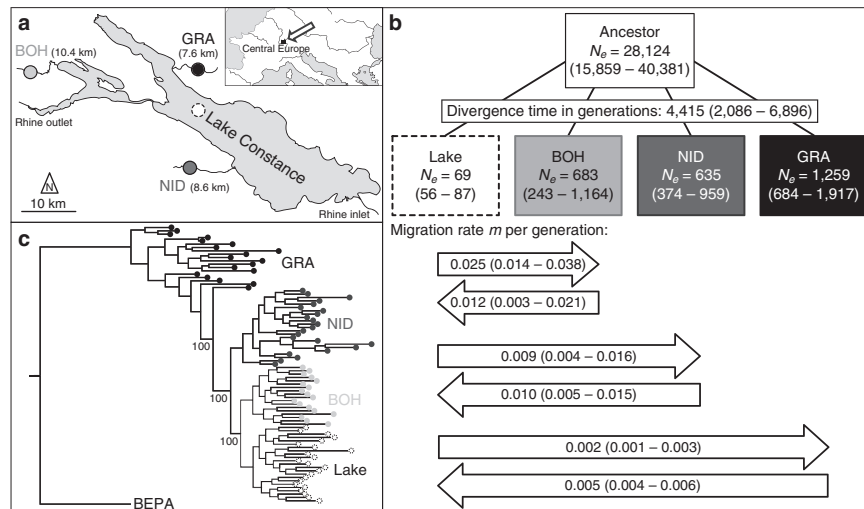


Figure 1 | Geographic context, demography and phylogeny of the study populations. (a) Location of the study populations from the Lake Constance basin, including the panmictic lake population and the three tributary stream populations (BOH, NID and GRA; the same colour coding and line types identifying the populations are used throughout the paper). Numbers in parentheses indicate the water distance between each stream site and the lake. (b) Estimated age of the split of the study populations from their common ancestor (divergence time), effective population sizes (numbers within boxes), and bi-directional migration rates between the lake and each stream population (numbers in horizontal arrows, representing the long-term proportion of immigration into the target population from the source population per generation forward in time). The values are based on an estimated SNP mutation rate of 6.8×10^{-8} . Numbers in parentheses are 95% bootstrap confidence intervals. (c) Phylogenetic relationship among the study populations visualized by a maximum likelihood tree rooted using a North American stickleback (BEPA, Bear Paw Lake, Alaska) as outgroup. Bootstrap support in per cent is given for the key nodes.

Next, we compared population-specific allele frequency spectra and found that across almost all minor allele frequency (MAF) classes, the lake exhibited the lowest and GRA the highest number of polymorphisms, with BOH and NID being intermediate (Supplementary Fig. 3). The lake also displayed the highest proportion of monomorphic SNPs, and the lowest proportion of tri-allelic SNPs (Supplementary Table 1). These findings clearly demonstrate that genetic diversity is lowest in the lake and increases from BOH to NID to GRA. Moreover, because the divergence among our study populations is recent (Fig. 1b, Supplementary Fig. 2) and the sharing of polymorphisms is extensive (Supplementary Table 1), most of the genetic variation in the present populations must have been standing in their common ancestor.

Calculating genome-wide baseline differentiation (that is, median F_{ST}) for each of the three lake–stream pairings revealed an increase in population differentiation from 0.005 in the lake–BOH comparison to 0.013 and 0.061 in the lake–NID and lake–GRA comparisons, whereas no stream–stream population comparison yielded baseline F_{ST} higher than 0.056 (BOH–GRA; NID–GRA: 0.047; BOH–NID: 0.012). In a rooted phylogeny, the lake population emerged as a distal branch nested within the more basal stream fish (Fig. 1c, Supplementary Fig. 4). An unrooted phylogeny further confirmed the close relatedness of the lake and BOH populations and the lower genetic diversity in the lake than in the streams (Supplementary Fig. 5).

Finally, we quantified linkage disequilibrium (LD) between all pairwise combinations of SNPs within all chromosomes in each population and found that strong allelic associations between SNPs occurred only over a scale of 1 kb or less; beyond this distance, LD was much weaker (Fig. 2a). The peak in LD at the smallest physical scale was driven by those SNPs exhibiting a high MAF; low-MAF SNPs exhibited more homogeneous and

generally weaker LD at all distances (Fig. 2a, insert). Another striking result was that the extent of LD across the genome was substantially greater in the lake population (and the two stream populations little divergent from the lake, that is, BOH and NID) than in GRA. A similar result was obtained by exploring average LD among marker pairs within non-overlapping chromosome windows: across most of the genome, LD was much stronger in the lake than in GRA (Fig. 2b), a result insensitive to the MAF threshold (Supplementary Fig. 6). Finally, the similarity in the local magnitude of linkage across the genome between the lake and each stream population, expressed as the correlation of LD between the chromosome windows, declined from the lake–BOH ($r = 0.17$) to the lake–NID ($r = 0.15$) and the lake–GRA pairing ($r = 0.12$) (all $P < 0.001$).

In combination, the above analyses resolve the demographic and selective history of stickleback in the Lake Constance basin. First, the demography is inconsistent with the view that the populations originate from a recent introduction of (presumably few) founder individuals, and instead supports an earlier postglacial and extensive natural colonization, presumably via the Danube drainage²³. Second, the demographic estimates of effective population size and all metrics of genetic variation make clear that the stream populations—and not the lake—represent the main reservoirs of genetic variation. This result is unexpected because Lake Constance is very large, and even conservative estimates of the present census size of its stickleback population range in the millions (personal communications from fishermen and fisheries authorities), which is certainly much greater than the size of any single stream population. (The streams investigated here are small, with an approximate average depth and width of 0.5 and 4 metres) Third, we observe the strongest genome-wide differentiation (F_{ST}) between a stream and the adjoining lake population, and not in any of the comparisons

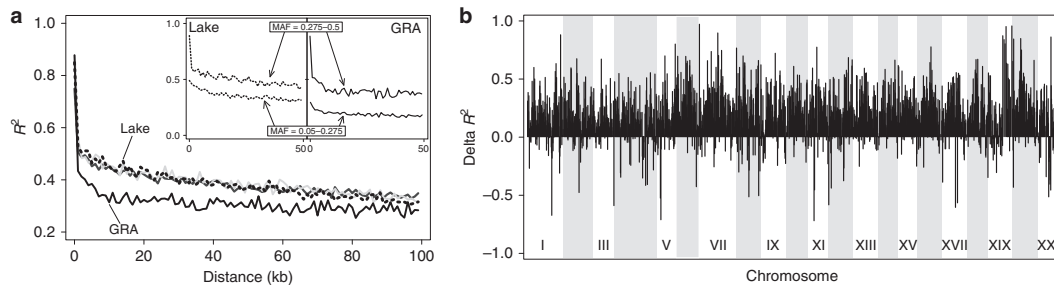


Figure 2 | Linkage disequilibrium across the stickleback genome. (a) Magnitude of LD (squared correlation of allele frequencies) between SNP pairs in relation to their distance on a chromosome, shown for the lake (dotted black line), BOH (solid light grey line), NID (solid dark grey line) and GRA (solid black line) population. The main panel uses a minimal MAF threshold of 0.05. The insert panels display LD separately for low-MAF (0.05–0.275) and high-MAF (0.275–0.5) SNPs in the lake and GRA population. (b) Difference in LD between the lake and GRA population along the genome. The data points represent the average LD in the lake minus the average LD in GRA across non-overlapping 200-kb chromosome windows, yielding a measure called Delta r^2 .

between the stream populations separated by dozens of kilometres of lake habitat. Fourth, the lake population proves to be phylogenetically derived from stream fish. All these observations can be brought in line by the biogeographically plausible perspective that the Lake Constance basin was initially colonized by ancestral stream-adapted stickleback. This colonization gave rise to multiple stream-resident populations isolated from each other by the adjoining, ecologically different lake habitat—that is, an ecological vicariance scenario. Subsequently, the lake fish started to adapt to their novel habitat and thereby experienced strong genome-wide selection. This selection should not only have reduced genetic variation in the lake relative to the streams, but also have driven relatively elevated LD within the lake, predictions clearly borne out by our analyses.

A key implication of this ecological vicariance scenario (visualized in Fig. 3d–f) is that the stream populations cannot be considered independent products of parallel divergence from an ancestral lake population. The stream fish are closer to the ancestral state while the lake population is the most derived. (Note that the phylogeny in Fig. 1c also rules out the possibility that the lake population results from a secondary colonization; in this case, the lake fish would branch basally from the stream populations.) Variation in the magnitude of genetic and phenotypic lake–stream divergence thus reflects different levels of homogenizing gene flow (that is, introgressive hybridization) from the large lake to the stream populations rather than variable progress in repeated parallel divergence (Fig. 3a–c). Supporting this view, typical lake phenotypes can sometimes be found at our BOH stream sample site during the breeding season (personal communication from fishermen). This highlights the potential for extensive genetic exchange in the one lake–stream pairing also exhibiting the highest migration rate estimates and the lowest genetic differentiation.

The strong genome-wide footprint of selection in the lake population, observed as relatively reduced genetic diversity and elevated LD, also raises an important methodological caveat. Marker-based approaches to demographic inference generally assume that allele frequencies reflect selectively neutral processes^{29–31}. In our study, the reduction of genetic variation by widespread selection in the lake clearly dissociates marker-based estimates of effective population size from biologically plausible census population sizes; the lake population, and to a lesser extent also the two stream populations strongly influenced by gene flow from the lake (BOH and NID), certainly have their estimated

effective population sizes biased downward relative to the GRA stream population. This highlights the benefit of backing up genetic inferences of demography with analyses of the selective history and with qualitative information from the field.

Genomically localized characterization of selection. The above genome-wide analyses indicated that the lake population has been particularly strongly influenced by widespread selective sweeps. To confirm this asymmetry in selection at a finer scale, we inspected localized signatures of selection at two classes of loci within the genome. The first, called F_{ST} extremes, included the 25 independent SNPs displaying the strongest lake–stream differentiation across all three lake–stream F_{ST} scans combined (79,770 total SNPs). None of these extreme SNPs showed fixed allelic differences between the habitats, but nearly so: F_{ST} ranged from 0.94 to 0.75—remarkably high values given the low baseline differentiation (Fig. 4a; genome-wide F_{ST} profiles visualizing the strikingly heterogeneous genomic divergence in all three lake–stream comparisons are provided in Supplementary Fig. 7). The F_{ST} extremes were found on 11 different chromosomes and derived mostly from the lake–GRA comparison that also produced the greatest baseline differentiation. Inspecting allele frequencies at the F_{ST} extremes showed that the MAF of the 25 SNPs were monomorphic) than in the corresponding stream population (with only four monomorphic SNPs; binomial test for similar occurrence of monomorphic SNPs: $P = 0.007$; Fig. 4a), suggesting that selection has mainly occurred, or has been more effective, in the lake. At the F_{ST} extremes, those alleles near fixation in one of the stream populations were generally also present in the other stream populations, with the frequency of these stream alleles increasing from BOH to NID to GRA (Fig. 4a). Finally, we found that haplotype decay around the F_{ST} extremes was slower in the lake than in the focal stream population (binomial $P = 0.004$; Fig. 4b).

The F_{ST} extremes represented genomic regions with nearly complete lake–stream allele frequency divergence, hence reflecting strong selection. To search for weaker or ongoing selective sweeps, we delimited a second class of loci based on haplotype structure^{32,33}. Specifically, we used R_{sb} ³⁴ to compare the rate of haplotype decay between the lake and the streams at 87,738 SNPs for each lake–stream comparison. Following the convention that positions with an absolute R_{sb} value > 4 provide compelling evidence of selection (for example, ref. 35), we

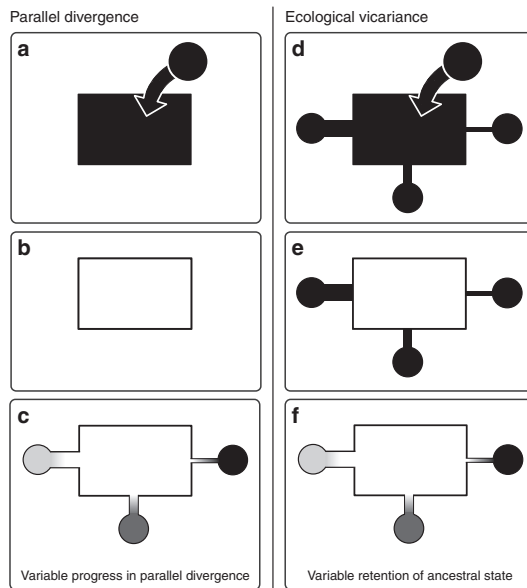


Figure 3 | Alternative demographic scenarios explaining repeated population divergence. The alternatives are exemplified by multiple stream populations divergent from the adjacent lake population in the Lake Constance basin. In the ‘parallel divergence’ scenario (panels a–c), a (stream-adapted) ancestor enters the lake (a) and becomes locally adapted (b). Subsequently, multiple stream populations derive independently via parallel evolution from the lake population (c), the latter thus representing their most recent common ancestor. The magnitude of lake–stream divergence in (c) (visualized as different grey shades) is determined by a combination of the time since colonization of each stream, the strength of local selection within each stream, and the extent of homogenizing gene flow from the lake into each stream. In this scenario, genetic variation available to local adaptation in the streams has been filtered during the adaptation of the lake population. Predictions here include greater genetic diversity in the lake than the stream populations, that F_{ST} is highest in stream–stream as opposed to lake–stream comparisons (due to founder events and relatively strong drift in these small populations), and that LD is highest in the streams (due to selective sweeps during adaptive divergence from the lake). In contrast, the ‘ecological vicariance’ scenario (panels d–f) involves the colonization of the entire study region by an already stream-adapted ancestor (d), followed by local adaptation in the lake (e). The magnitude of lake–stream divergence is then primarily determined by the extent to which the stream populations can maintain their genetic integrity in the face of gene flow from the large lake population (f). Predictions here include greater genetic diversity in the streams than in the lake, highest F_{ST} in lake–stream as opposed to stream–stream comparisons, and strongest LD in the lake due to extensive selection. All these latter predictions are confirmed by our analyses.

identified a total of 22 such ‘Rsb extremes’ on 11 chromosomes across all three lake–stream comparisons (lake–stream Rsb profiles are presented as Supplementary Fig. 8; in contrast to the F_{ST} extremes, Rsb extremes emerged from all lake–stream contrasts, Supplementary Fig. 9). Interestingly, examining allele-specific haplotype structure revealed that within both habitats, the lake alleles were surrounded by relatively longer haplotype tracts than the alternative stream alleles (Fig. 4c). This indicates that alleles selected positively in the lake, but presumably negatively in the streams, are maintained at

substantial frequency in the streams by gene flow from the lake population. Finally, our haplotype-based analysis also revealed signatures of selective sweeps that have occurred in the stream habitat (Fig. 4d, Supplementary Fig. 9).

Overall, our analyses of localized signatures of selection provide strong support for the selective scenario indicated by the genome-wide signatures: selection is wide-spread across the genome and is asymmetric, with more extensive sweeps having occurred in the lake than in the stream populations. Moreover, lake–stream divergence in the Lake Constance basin has clearly occurred in the face of gene flow. Consistent with the census size (but not the estimated effective population size) of the Lake Constance population being orders of magnitude larger than the stream populations, introgression occurs primarily from the lake into the streams. Nevertheless, many loci resist gene flow and maintain substantial differentiation from the lake^{12,15–17}, thereby generating heterogeneous genomic divergence between the lake and the stream populations³⁶.

Signatures of selection around a known adaptation locus. Our analyses of localized signatures of selection within the genome focused on regions likely important to adaptation to the lake and stream habitats, yet it is unknown what phenotypes the polymorphisms in these regions influence. For the extent of lateral plating, however, it was possible to take an alternative route and to investigate the molecular signatures produced by selection on a trait known *a priori* to be important to lake–stream divergence. We started at the phenotypic level by establishing that lake individuals were mostly completely plated, whereas plating was relatively reduced in all stream populations, most clearly so in NID and GRA (lake–BOH permutation test for similar plating: $P = 0.420$; lake–NID: $P = 0.002$; lake–GRA: $P < 0.001$) (Fig. 5a). This agrees with earlier work using different populations and/or samples from the same basin^{21,23}. Next, we performed a bulk segregant analysis (BSA) by pooling all completely and all low-plated stream fish into two separate groups. Genetic differentiation between these groups across genome-wide SNPs revealed a region on chromosome four (ChrIV) harbouring markers with a very strong association between allelic state and phenotype (Fig. 5b). The peak association ($F_{ST} = 0.78$) occurred immediately downstream of the Ectodysplasin (*Eda*) gene. This locus is known as major determinant of lateral plating^{37,38}, with a causative *cis*-regulatory polymorphism having been identified 1 kb downstream of the coding region³⁹. No SNP outside this region on ChrIV displayed $F_{ST} > 0.38$.

Combined, the phenotypic data and BSA indicate that differentiation in plating among our study populations involved adaptive lake–stream divergence at the *Eda* locus. We thus predicted molecular footprints of selection at this locus. To evaluate this prediction, we inspected all three lake–stream F_{ST} scans for the magnitude of differentiation around *Eda* (Fig. 5c). As expected from the plate morph distribution (Fig. 5a), the strongest differentiation occurred in the lake–GRA comparison ($F_{ST} = 0.40$), just 5.7 kb downstream of *Eda*. However, in this particular comparison, the most divergent SNP near *Eda* ranked only within the upper 3.5 percentile of the genome-wide F_{ST} distribution (lake–BOH and lake–NID comparisons: 8.8 and 2.3 percentile). Similarly, the highest absolute Rsb value around *Eda* (1.17) also emerged from the lake–GRA comparison but fell only within the upper 23 percentile of the genome-wide Rsb distribution. Hence, thousands of SNPs displayed a stronger deviation from selective neutrality than the *Eda* locus. Accordingly, subjecting the lake–GRA pairing to a standard selection outlier detection analysis (*BayeScan*⁴⁰) failed to provide any evidence of selection at SNPs surrounding *Eda* (Supplementary Fig. 10).

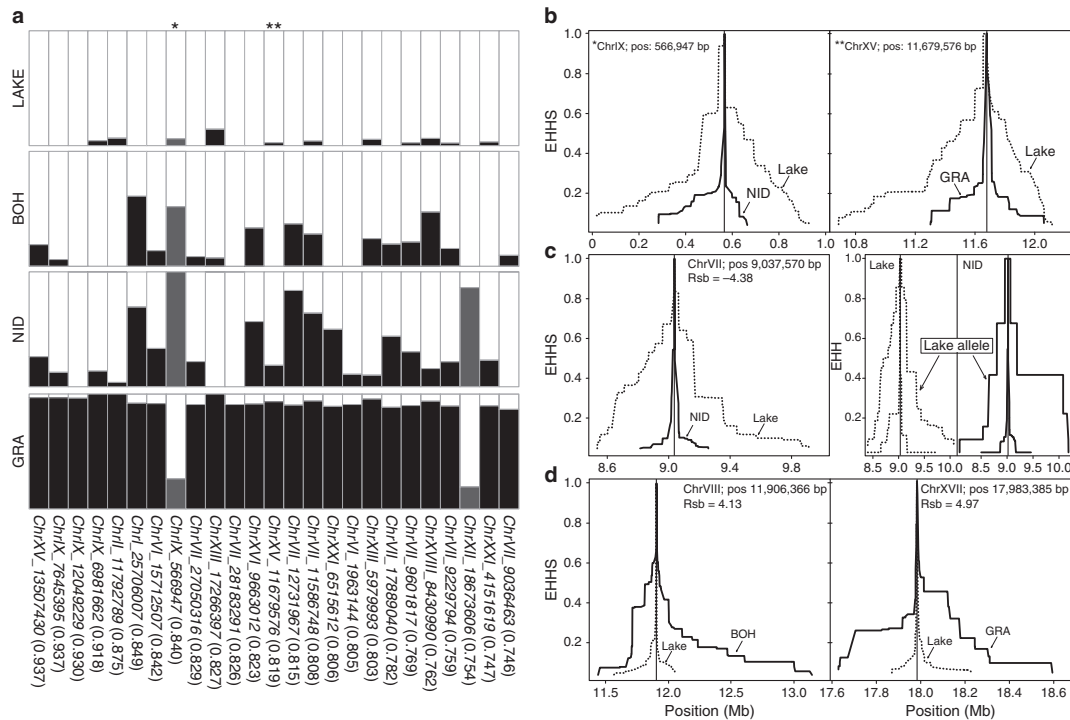


Figure 4 | Localized signatures of selection. (a) Allele frequencies within each population at the 25 F_{ST} extremes (columns). The proportion of the SNP alleles predominant in the lake are shown in white, while the proportion of the alleles predominant in the streams are either black (when the extreme F_{ST} value emerged from the lake-GRa genome scan, $N = 23$) or dark grey (extreme F_{ST} value observed in the lake-NID scan, $N = 2$). On the bottom, the genomic position and the highest F_{ST} value observed across all lake-stream comparisons are given for each F_{ST} extreme. **(b)** Lake and stream haplotype decay (EHHS) around representative F_{ST} extremes (flagged by asterisks in **a**) identified in the lake-NID F_{ST} scan (left panel) and in the lake-GRa scan (right panel). **(c)** Haplotype decay in the lake and the NID stream population around a representative negative Rsb extreme identified in the lake-NID Rsb scan (left panel). For the same Rsb extreme, the right panel displays allele-specific haplotype decay (EHH) around each allele within each population separately (both alleles occur in both populations; the allele predominant in the lake is labelled 'Lake allele'). **(d)** Lake and stream haplotype decay around two representative positive Rsb extremes identified in the lake-BOH Rsb scan (left panel) and in the lake-GRa scan (right panel). Note that the scale of the x axis varies in **b-d**.

More nuanced insights into the evolution of lateral plating were obtained by analysing haplotype structure around the *Eda* locus: in the streams, where both *Eda* alleles (still) occur at substantial frequencies, haplotype decay was slower around the allele associated with complete plating (Fig. 5d, top). Moreover, the haplotype structure around the completely plated allele in the streams matched the haplotype structure around this allele in the lake (where the low-plated allele was too rare to characterize LD) (Fig. 5d, bottom). Together, this indicates that selection for complete plating in the lake has been more effective than selection against plates in the streams, and again suggests the maintenance of an unfavourable variant—and the associated phenotype—in the streams by gene flow from the lake (see also Fig. 4c). To fully appreciate the extent of LD driven by selection on lateral plating, we again took a bulk segregant approach by treating all completely and low-plated stream fish as separate groups, and looked for distortions between these groups in the rate of haplotype decay along ChrIV. This confirmed that selection on the *Eda* variant driving complete plating has been much more intense than selection on the low-plated variant, and showed that the associated sweep has influenced haplotype structure at the scale of megabases (Fig. 5e). Unexpectedly, this

scan also detected a second, similarly strong selective sweep in completely plated stickleback centred at 11.4 Mb. This latter region also exhibited a clear signature of divergence in the F_{ST} -based BSA (Fig. 5b, top): the differentiation peak in this region ($F_{ST} = 0.31$) fell within the top 0.06 per cent of the genome-wide distribution.

Together, the investigations at the *Eda* locus highlight our limited ability to elucidate the genetic basis of adaptive population divergence based on genetic markers when selective sweeps are incomplete. Neither the magnitude of differentiation (F_{ST}) nor haplotype structure (Rsb) among populations allowed the major plate locus to emerge as an obvious selection candidate—despite substantial evolution in the associated ecologically important phenotype, and despite an extensive selective sweep visible when comparing haplotype structure among individuals grouped by phenotype. Given that stronger signatures than those around *Eda* are numerous in our data sets, we conclude that hundreds of genomic regions must be involved in the adaptive divergence into lake and stream habitats. We further propose that lateral plate evolution in the Lake Constance basin is governed by at least one other locus besides *Eda*. Inspecting the newly detected region on ChrIV indeed produces a

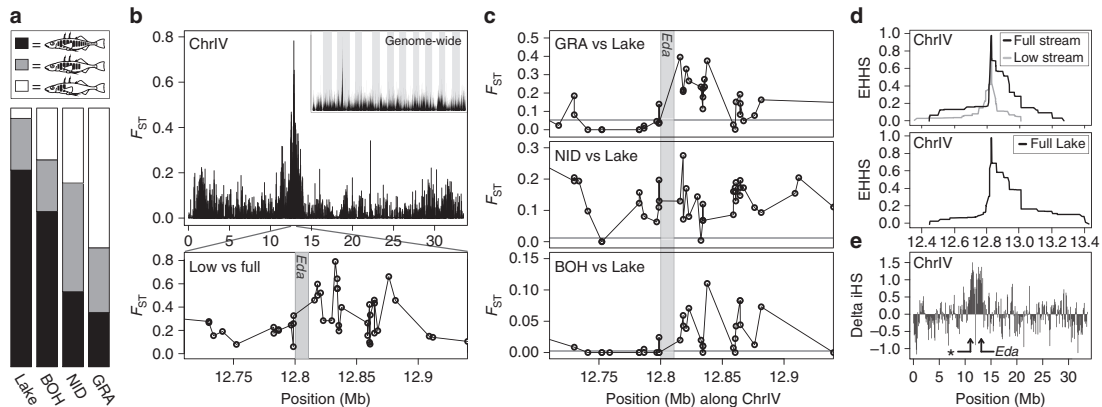


Figure 5 | Lake-stream divergence in lateral plating and the associated molecular signatures. (a) Frequency of the three plate phenotypes (completely, partially and low-plated) in the four study populations. (b) Genetic differentiation (F_{ST}) between completely and low-plated stream stickleback reveals a peak on chromosome IV, with the highest F_{ST} value immediately downstream of the *Eda* coding region (grey vertical bar). (c) Divergence (F_{ST}) profiles around the plate locus for all lake-stream comparisons, with the horizontal grey lines indicating baseline differentiation (note that the scale of the y axes varies). (d) Haplotype decay (EHHS) in completely and low-plated stream stickleback around the *Eda*-associated SNP exhibiting the strongest phenotype-genotype association in the bulk segregant analysis (see b, bottom) (upper panel). The lower panel shows haplotype decay around the completely plated allele at the same SNP in the lake population. (e) Profile of the difference in the rate of haplotype decay between completely and low-plated stream stickleback along chromosome IV. High positive values of this 'Delta iHS' metric, indicating more extensive haplotype tracts in completely plated individuals, occur across a broad region centred at the *Eda* locus, and in a second genomic region nearby (indicated by an asterisk; note the corresponding F_{ST} peak in the upper panel of b).

strong candidate gene, *Col23a1* (bp-position 11,443,468–11,468,190; this specific segment contained the highest- F_{ST} SNP observed across the new candidate region in the BSA). Like *Eda*, this gene encodes a transmembrane collagen involved in the development of the epidermis⁴¹. Since the new candidate region and *Eda* occur in close proximity (c. 1.4 Mb apart) in a low-recombination chromosome region⁴², it is tempting to speculate that the coupling of alleles in the two regions might facilitate divergence in plating relative to the situation where each locus segregates independently^{12,43}.

Detection and characterization of inversions. Our genetic data indicate that lake-stream divergence in the Lake Constance basin has occurred in the face of gene flow. Genetic polymorphisms predicted by theory to resist homogenizing gene flow and to diverge between populations particularly well are chromosomal inversions^{43–45}. The reason is that different inversion types can physically couple alleles promoting adaptation to different habitats across multiple loci. The integrity of these allele clusters is easily maintained, because a single crossover within the inversion generally produces unbalanced meiotic products in inversion heterozygotes (that is, heterokaryotypes), thus effectively suppressing recombination^{46,47}. Consequently, alternative inversion types can be considered single large-effect alleles.

To test this idea, we examined if lake-stream divergence in the Lake Constance basin was promoted by chromosomal inversions. For this, we scanned the genome for extended distortions in the relative RAD sequence coverage between the lake and each stream population (Supplementary Fig. 11). This produced three strong candidates, located on ChrI (approximate length: 500 kb), ChrXI (450 kb) and ChrXXI (2.1 Mb) (Fig. 6a)—all coinciding with inversions recently identified in a comparison of marine and freshwater stickleback¹¹. For two of these candidate inversions (ChrI and ChrXI), we designed PCR primers across expected inversion breakpoints based on our RAD sequences, and the presence/absence of PCR products confirmed that these regions

were inversions (Supplementary Fig. 12). We then performed several complementary analyses to characterize the three inversion polymorphisms in our populations. Inspecting inversion-specific allele frequencies revealed that the lake population was consistently fixed for one inversion type, whereas the stream populations were polymorphic at two (NID) or all three inversions (BOH and GRA). However, only at the ChrI inversion were lake-stream frequency shifts strong enough to drive clearly elevated F_{ST} relative to baseline differentiation (Fig. 6b). Consistent with only the stream populations being polymorphic for the inversions, the allelic diversity at polymorphic sites within the inversions tended to be elevated in the stream populations relative to the lake (Fig. 6c). However, the segregation of an inversion type at very low frequency within a population sometimes generated an excess of SNPs displaying reduced diversity relative to the genomic baseline within that population (BOH and NID at the ChrXI and ChrXXI inversions, Fig. 6c). The stream populations also exhibited a clear excess of SNPs falling into the specific MAF class mirroring the relative frequency of the minor inversion type (Fig. 6d). SNPs within this MAF class—but not those from other MAF classes—revealed extended blocks of nearly perfect LD caused by the inversion polymorphisms in the streams (Fig. 6e).

For the ChrI inversion, we experimentally confirmed suppressed recombination in inversion heterozygotes by inspecting crossover frequencies in an F2 intercross derived from two parental individuals homozygous for either inversion type^{38,42}. Not a single crossover occurred within the inversion, but recombination immediately adjacent to the inversion was frequent (Fig. 6f; see Supplementary Fig. 13 for a negative control of this analysis). Nevertheless, for large inversions, theory predicts that occasional double crossovers should allow some genetic exchange between the inversion types, albeit not near the inversion breakpoints^{47,48}. We examined this prediction for the ChrI inversion by comparing homozygotes for one inversion type to homozygotes for the other type, considering individuals from

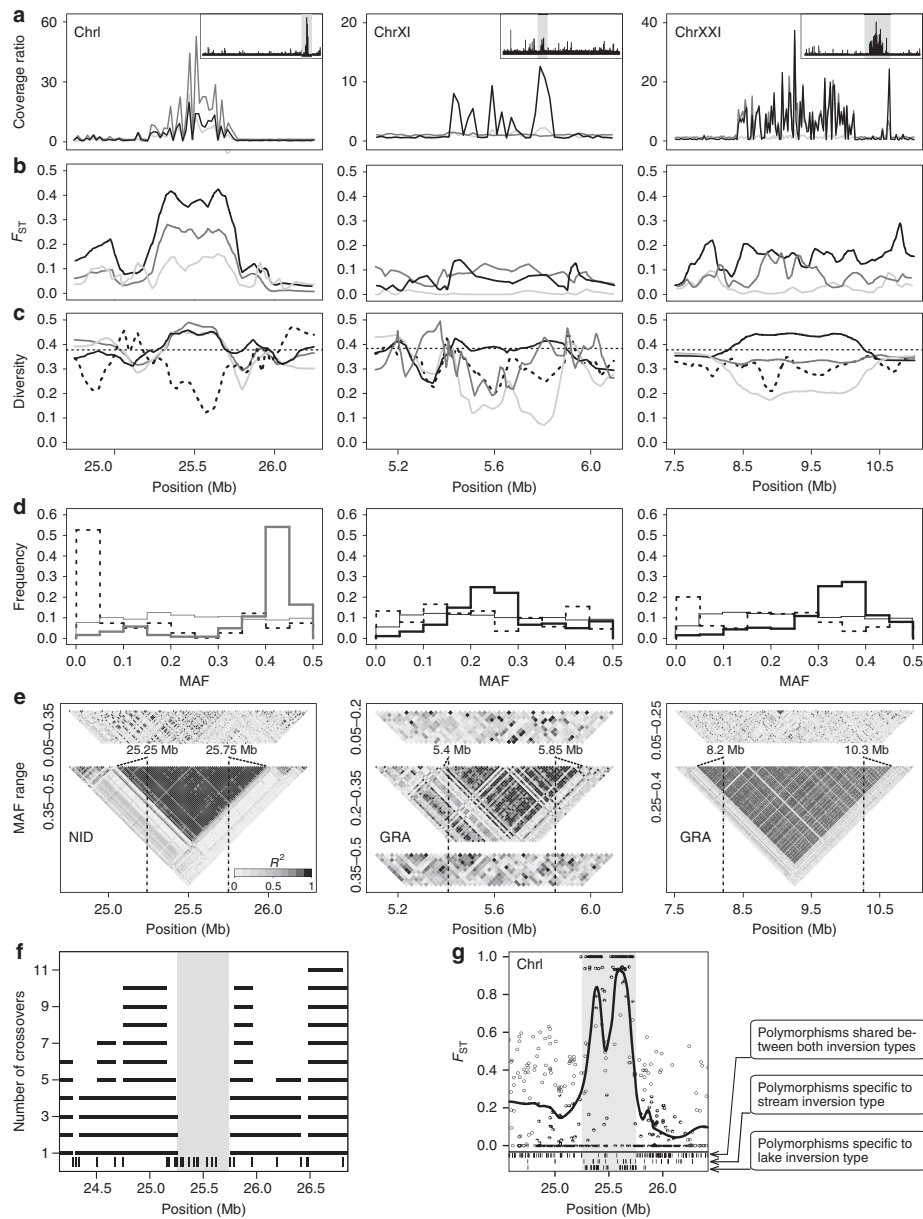


Figure 6 | Detecting and characterizing large chromosomal inversions in lake and stream stickleback. (a) RAD sequence coverage of the stream populations relative to the lake population (coverage ratio), shown separately for each lake–stream population pair (lake–BOH light grey, lake–NID dark grey and lake–GRA black) along focal segments of the chromosomes I, XI and XXI harbouring an inversion. The inserts show the coverage ratio along the entire chromosomes (focal segments shaded grey), based on the coverage data pooled across the three stream populations. (b) Genetic differentiation between the lake and each stream population, and (c) allelic diversity at SNPs within each of the four populations, around the three inversions. (d) MAF distribution for SNPs located within each inversion, shown for the stream population exhibiting the strongest coverage ratio distortion relative to the lake (see a) (thick solid line). For comparison, the genome-wide MAF distribution for the same stream population (thin solid line), and the MAF distribution for all inversion SNPs in the lake population (dotted line) are also plotted. (e) Linkage disequilibrium heat maps based on SNPs from distinct MAF classes shown for the same stream populations as in d. (f) Recombination between the ChrI inversion types (the inverted segment is shaded grey) in a laboratory cross. Black horizontal bars indicate the number of crossovers observed between two neighbouring markers (vertical bars on the bottom), and the grey profile shows the corresponding recombination rate. (g) Genetic differentiation (raw values and smoothed profile) between individual pools of ChrI inversion homozygotes. The bottom part indicates the position of SNPs shared between the two inversion types, and of those unique to each.

all populations. We found that while these two groups were fixed for different SNP alleles across most of the inversion, differentiation decayed in a narrow region in the centre of the inversion (Fig. 6g). This region was also relatively enriched for polymorphisms shared between the two inversion types, but contained relatively few SNPs unique to either of the two types (Fig. 6g, bottom).

To learn more about the history and ecology of the three inversion polymorphisms, we next established the phylogenetic relationship among our study individuals using haplotype information based on SNPs located within the inverted regions only. For each inversion, this revealed the presence of two haplotype clusters separated by a deep split (Fig. 7a). In line with our findings from the allele frequencies at putative loci under selection (Fig. 4a), Lake Constance fish consistently harboured haplotypes from one of these clusters only, whereas all stream populations contained haplotypes from both clusters. Repeating the phylogenetic analysis by including SNPs extracted from 21 previously sequenced marine and freshwater stickleback sampled across the species' global distribution¹¹ produced a striking result: haplotypes representing the inversion type for which the Lake Constance population was fixed clustered consistently or were even identical with haplotypes recovered in marine stickleback (Fig. 7b). Conversely, haplotypes representing the inversion type found exclusively in the streams were closely related to, or identical with, haplotypes from global freshwater populations. To further explore how consistently these inversion polymorphisms are recruited for lake–stream divergence, we investigated SNP data for individuals sampled from Lake Geneva and from one of its tributary streams, waters documented to have been colonized by stickleback very recently (nineteenth century) and independently from the Lake Constance basin (see references in refs 21,22; genome-wide divergence in this lake–stream pair is described in Supplementary Fig. 14). We

here again recovered all three inversion polymorphisms (Fig. 7c, Supplementary Fig. 14). At the ChrI inversion, the direction of lake–stream divergence was congruent between the Lake Constance and Lake Geneva basins, whereas the ChrXI showed no divergence in the latter. Surprisingly, the direction of lake–stream divergence at the ChrXXI inversion was reversed between the two basins.

Overall, a first insight emerging from our analyses of inversions is that the relative frequencies of inversion types need to be taken into account when scanning population genomic data for the presence of such polymorphisms. Characteristic signatures like extended blocks of SNPs displaying exceptional levels of population differentiation or strong LD can become evident only when restricting SNPs to the appropriate MAF class. Second, our analysis of the ChrI inversion shows that genetic exchange between inversion types can occur despite effective overall recombination suppression, and that this exchange is biased towards the inversion centre. To our knowledge, this has previously been demonstrated only for much larger inversions in *Drosophila* and *Anopheles*^{49,50}. Our data from the laboratory cross further suggest exceptionally high recombination rates in the collinear segments immediately flanking the inversion (Fig. 6f). This is unexpected—double crossover encompassing a single inversion breakpoint should produce unbalanced chromatids, hence one would predict relatively reduced recombination in these regions⁴⁷.

Finally, the distribution of inversion haplotypes in the Lake Constance basin suggests divergent lake–stream selection on these chromosomal rearrangements. Specifically, the occurrence of shared haplotypes at both inversion types within multiple, presently unconnected stream populations, and the consistent presence of only a single inversion type in the lake, indicate particularly effective sorting of ancestral standing variation in the lake population. This reinforces our conclusion of asymmetric

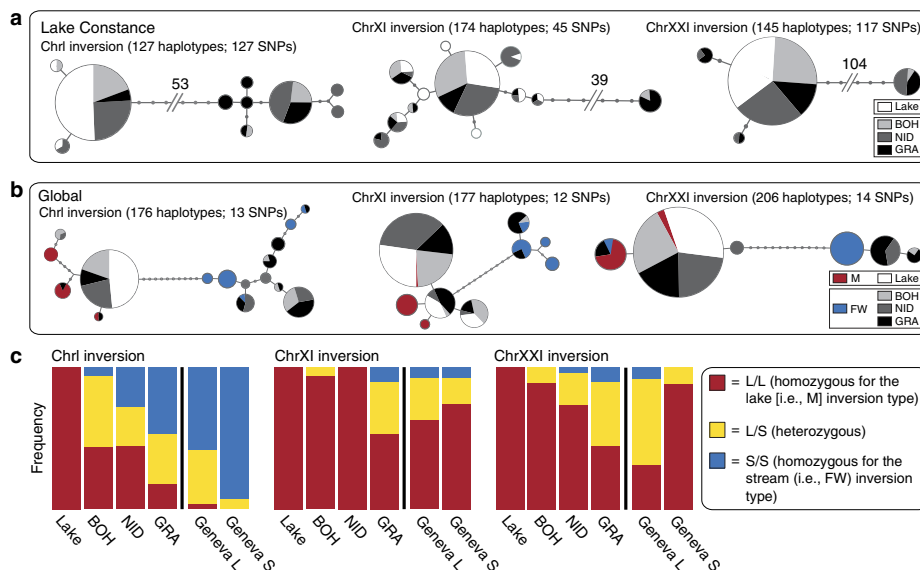


Figure 7 | Phylogenies and habitat associations of the inversions. (a) Haplotype phylogenies restricted to SNPs located within the inversions, based on individuals from the Lake Constance basin. Pie sizes reflect the relative frequency of each haplotype, and internodes are mutational steps. Only haplotypes recovered more than twice are shown (total haplotype numbers are given above the networks). (b) Haplotype networks as in a, but additionally including individuals from marine and freshwater populations across the species' global range¹¹. (c) Frequency of the three diploid genotype classes at each inversion in the four populations from the Lake Constance basin, and in the lake–stream pair from the Lake Geneva basin. The colour coding follows b.

selection based on the genome-wide analyses and the inspection of F_{ST} and R_{sb} extremes, and supports the view that inversion polymorphisms are ecologically relevant⁴⁴. (We note that we did not find any indication of intrinsic incompatibility or transmission disequilibrium between the inversion types, as their frequencies did not deviate from Hardy–Weinberg expectation in any inversion–population combination. Details not presented, but see Fig. 7c.)

All inversion haplotypes occurring within Lake Constance either coincide with haplotypes predominant in marine stickleback. This suggests the presence of shared selective features between the ocean and large lakes—possibly mediated by a pelagic lifestyle in both habitats (see ref. 51 for similar evidence in trout)—driving genuine parallel evolution at a much larger geographic scale than our focal lake–stream system. In any case, since inversions are not (only) relevant to saltwater–freshwater adaptation¹¹. To further complicate functional conclusions, the ChrXXI inversion has diverged in opposed directions between lake and stream stickleback in the Lake Constance and the Lake Geneva systems. This unexpected trend is unlikely to arise from drift in the young Geneva system: among the 50 most extreme genome-wide F_{ST} values in this exceptionally weakly divergent lake–stream pair (genome-wide median $F_{ST} = 0$), 22 (44%) map to the ChrXXI inversion, including the top value observed overall ($r = 0.338$) (Supplementary Fig. 14). This suggests intense selection on this inversion polymorphism in the Geneva system. However, given the great number of genes coupled by each inversion (~24, 25 and 109 genes for the ChrI, ChrXI and ChrXXI inversions), dissecting the precise target(s) of selection in different ecological contexts will remain a serious challenge. Finally, the detected sharing of haplotypes between our study populations (derived from Atlantic ancestors) and worldwide stickleback populations (including Pacific-derived fish), along with the vast mutational differentiation observed between the inversion types (Fig. 6g and Fig. 7b), indicates that all three inversion polymorphisms must be ancient.

To summarize, a main goal of our study was to dissect the geographic and selective history of adaptive diversification in lake and stream stickleback populations within a single lake system. Combining demographic inference with broad scale and fine-scale analyses of genetic differentiation and diversity, linkage disequilibrium and haplotype structure within the genome allows us to reject a standard scenario of parallel divergence of multiple stream populations from a shared ancestral lake population (Fig. 3a–c). Instead, our results support a history of ecological divergence with gene flow. This latter scenario involves the widespread colonization of the Lake Constance basin by a lake-adapted ancestor, the subsequent emergence of a derived stream-adapted population through intense selection of standing variation and sustained gene flow across the lake–stream boundaries (Fig. 3d–f). Consequently, different magnitudes of overall divergence among the lake–stream pairings, and heterogeneous lake–stream divergence across the genome, do not mirror how strongly gene flow from the lake has constrained emergence of adaptation in the streams, but how effectively migration from the lake has eroded initial stream adaptation. Our work thus underscores that investigations of patterns of divergence consistent with parallel evolution should consider an alternative—that is, the repeated retention of shared ancestral variation, and should be rooted in detailed knowledge about the geographic and selective history of populations¹⁴. Nevertheless, rooted within a vicariance background, our investigation of inversion polymorphisms indicates the recycling of the same genetic variants for adaptive divergence in seemingly different ecological contexts, and hence real parallel evolution on a large geographic scale.

Furthermore, our finding of highly heterogeneous genomic divergence conflicts with the recent theoretical prediction that adaptive divergence in the face of gene flow involving selection on extensive standing variation should produce genome-wide reproductive isolation and therefore limit heterogeneity in genome divergence⁵². Given the numerous factors influencing adaptive divergence in natural populations, we believe that it will remain very difficult to predict how fast and to what extent heterogeneous genomic divergence should build up. However, our study clearly supports the notion that heterogeneity in genome divergence is promoted by sustained gene flow between young populations adapting to ecologically different environments (for example, ref. 53). We challenge the claim that such heterogeneity represents the divergence of populations after reproductive isolation has become complete⁵⁴.

Finally, our study adds molecular evidence to the idea that chromosomal inversions promote adaptive divergence by acting as loci of large effect⁴⁴. However, lake–stream stickleback divergence certainly also involves numerous loci not located within chromosomal rearrangements, and selection on some of these loci appears at least as strong as selection on the inversions. Determining the importance of inversions relative to other adaptive polymorphisms in evolutionary diversification remains an important empirical issue.

Methods

Stickleback samples and marker generation. Specimens from the Lake Constance population were sampled at two localities (Romanshorn, Switzerland, $N = 12$, and Unteruhldingen, Germany, $N = 13$; for geographic details see ref. 23). Genetic structure is absent at any scale within this large lake (Supplementary Fig. 1 and refs 21,23), so the two lake samples were combined to a single ‘lake’ pool for all analyses. Stream stickleback were sampled from three geographically well-separated tributaries connected through the lake only (Fig. 1a). The stream sites correspond to the Bohlingen (BOH, $N = 22$), Nideraach (NID, $N = 24$) and Grasbeuren (GRA, $N = 24$) localities in ref. 23 (for details on all specimens see Supplementary Table 2). Natural dispersal barriers are absent in all streams, but low man-made dams have likely restricted gene flow from the lake to the NID and GRA sites over the last decades. All work in this study was approved by the Veterinary Office of the Canton of Basel-Stadt (permit number: 2383).

DNA was extracted from stickleback fin and muscle tissue using either a MagNA Pure LC278 extraction robot (Roche, Basel, Switzerland) with the tissue Isolation Kit II, or the DNeasy Blood & Tissue Kit (Qiagen, Valencia, USA). After an RNase treatment, the extracts were standardized to $18 \text{ ng } \mu\text{l}^{-1}$ based on multiple NanoDrop photospectrometer readings (Thermo Scientific, Wilmington, USA), and used to generate RAD DNA libraries essentially following the protocol described in ref. 5. The main modification was that we used the *Nsi*I enzyme for DNA restriction, exhibiting a 7.5 times higher recognition site density (that is, c. 164,000 sites across the 460-Mb stickleback genome) compared with the commonly used *Sbf*I restriction enzyme. We prepared 12 total RAD libraries, each combining individually 5mer-barcoded DNA from seven or eight of the 95 total individuals. For final enrichment, we pooled six replicate PCRs per library to reduce amplification bias.

Each library was single-end sequenced with 100 cycles on a separate Illumina HighSeq2000 lane. Raw sequence reads were parsed by individual barcodes and aligned to the improved assembly⁴² of the threespine stickleback reference genome¹¹ by using Novoalign v2.07.06 (<http://www.novocraft.com/products/novoalign>; sequencing and alignment statistics are provided in Supplementary Table 2). We enforced unique alignment, tolerating an equivalent of ~8 high-quality mismatches or gaps (flags: -t236, -g40, -x15). Alignments were BAM-converted in Samtools v0.1.11 (ref. 55). For individual consensus genotyping, we first applied two effective filters to further exclude RAD loci located on repeated elements. First, loci were excluded if they displayed a read coverage exceeding three times the mean coverage across all loci within an individual. Second, if a RAD locus was polymorphic, it was excluded if the two dominant haplotypes failed to account for >70% of all reads.

Loci passing the above filters were subjected to consensus genotyping using a refinement of our earlier haplotype-based algorithm⁵, which has been demonstrated to perform highly accurately²⁶. The main novelty was that instead of building genotypes quality-aware base-by-base, we discarded sequence quality and treated the entire read as the genotyping unit. A diploid genotype was called if the read coverage contributed by the two dominant haplotypes, or the total coverage for monomorphic loci (‘effective coverage’), was 15 or greater (median total coverage across all RAD loci and individuals was $38.5 \times$). Because we observed in our previous work that the distribution of the two haplotypes for heterozygous loci was over-dispersed relative to the binomial expectation, we avoided distinguishing

homozygote from heterozygote genotypes based on a theoretical distribution. Instead, a locus was considered heterozygous if the ratio of the second most frequent haplotype to the sum of the first and second was > 0.25 . Otherwise, a locus was considered homozygous. If the effective coverage was below 15 but at least two, we called a haploid genotype only, based on the dominant haplotype. Loci with single-read coverage were discarded. Inspection of the haplotype distribution at RAD loci showed that with our sequence data, this defensive algorithm maximized both the detection of truly heterozygote loci and the exclusion of polymorphisms reflecting technical artifacts (Supplementary Fig. 15). To create the raw SNP matrix for downstream analyses, we pooled the consensus genotypes across all populations and extracted a maximum of six SNPs per RAD site, provided the haploid consensus genotype coverage across all individuals and populations was at least $80\times$.

Demography and phylogenetics. To explore the evolutionary history of our four study populations, we reconstructed their demography using the coalescent simulator *fastsimcoal2.1* (ref. 29). As input, we computed the observed joint site frequency spectrum (SFS) for each of the six pairwise population combinations. For this, we first sampled at random exactly 30 haploid consensus genotypes per RAD locus from each population. Loci with sparser coverage and those harbouring more than two polymorphisms with an identical frequency of the less common allele (that is, the 'minor allele frequency' (MAF)) across the last 30 positions were ignored. The latter excluded uninformative sequential pseudo-SNPs from RAD loci harbouring a micro-indel polymorphism, and hence ensured that only true SNPs were considered. Next, we counted the occurrence of the minor allele at each of the 89 positions per RAD locus in each population to populate the SFS. This considered both monomorphic positions and bi-allelic SNPs. For the latter, the minor allele was defined based on the pool of all four populations. If the MAF of a SNP was exactly 0.5, both alleles were treated as minor and entered the SFS, but with a weight of 0.5 only (personal recommendation by L. Excoffier). The resulting joint SFS were based on 14,837 million base positions on 166,711 RAD loci. We additionally computed all population-specific SFS with the same resolution.

Using the observed joint SFS, we then performed simulations with *fastsimcoal2.1* to estimate the most likely parameter values for an evolutionary scenario in which the four focal populations split under gene flow from an ancestral population colonizing the Lake Constance basin. We here assumed that the populations in the different habitats established rapidly, justifying a single splitting time. We estimated the age of the split, all effective population sizes (including the ancestor), migration rates between the lake and each stream population (but not among stream populations) and the SNP mutation rate. The simulation was run in 80 replicates, each including 40 estimation loops with 100,000 coalescent simulations. To determine the best parameter estimates, we selected the 10 most likely replicate runs (that is, those with the smallest difference between the estimated and observed likelihood) and used this subset to calculate the mean for all parameters, along with their 95% confidence intervals (95 percentiles from bootstrap distributions based on 100,000 resamples). Because the lake population turned out to be particularly strongly influenced by selection, we explored an analogous model in which just the two stream populations most divergent from the lake (NID and GRA) split from an ancestor under gene flow. The joint SFS, the simulation template files, the parameter estimation files, and the command line settings used to run *fastsimcoal2.1* are provided as Supplementary Data 1–12.

To explore phylogenetic relationships among populations, we first reduced individual genotypes to single-letter code and eliminated individuals with $> 75\%$ and SNPs with $> 15\%$ missing data. We then used the R (ref. 57) package *phangorn* (ref. 58) to infer the most appropriate model of sequence evolution⁵⁹ (GTR + G + I). (The R language was used for all analytical procedures in this paper, unless noted otherwise.) Finally, we constructed unrooted maximum likelihood trees to infer the phylogeny of all four populations (based on 51,188 SNPs) and of the two lake samples only (55,561 SNPs). These analyses used no more than one SNP per RAD locus and required a MAF > 0.2 across all populations (MAF > 0.05 resulted in very similar results). Node support was assessed with 200 bootstrap replicates. The same data were also used to visualize genetic structure based on a principal coordinates analysis as implemented in the R package *ape* (ref. 60). Rooted phylogenies were constructed analogously by incorporating genotype data from geographically distant outgroup stickleback individuals, including the Pacific BEPA reference genome individual, at 14,429 SNPs ascertained in the populations from the Lake Constance basin.

Genetic diversity. Two analyses were conducted to compare genetic diversity among the populations. For both, we only considered SNPs from our raw SNP matrix that occurred alone on a given RAD locus (that is, data from RAD loci harbouring multiple polymorphisms were ignored). Using only such 'loner SNPs' avoided potential bias in the estimation of genetic diversity due to pseudo-SNPs caused by micro-indels. We further ignored those loner SNPs displaying a minor allele count < 2 across all individuals pooled, thereby avoiding sequencing artifacts. We thus obtained a total of 62,332 genome-wide loner SNPs. As a first measure of diversity, we determined for each population the proportion of the total loner SNPs actually being polymorphic. To obtain a second diversity measure, we screened all loner SNPs for the presence of three alleles across all individuals pooled ('tri-allelic

loner SNPs'; the least frequent allele had to occur at least twice across all individuals). On average, one out of 169 loner SNPs proved tri-allelic (genome-wide total: 368). We then determined for each population the proportion of the total tri-allelic loner SNPs actually displaying all three alleles.

Genome-wide LD. We quantified LD within each population using the squared correlation coefficient (R^2) between pairs of SNPs. From the raw SNP matrix, we excluded SNPs that were tri-allelic or had $> 25\%$ missing genotypes, and individuals with $> 75\%$ missing diploid genotype calls. The remaining SNPs were filtered for two different MAF ranges (0.05–0.275 and 0.275–0.5). Only a single randomly chosen SNP was retained if multiple SNPs passed these thresholds for a pair of sister RAD loci (that is, the two RAD loci flanking the same restriction site). The final number of SNPs was 16,088 and 18,787 for the former and latter MAF range (marker number was adjusted to be equal for all populations). We then ran PLINK (ref. 61) with the command line '-ld-window 100 -ld-window-kb 100 -ld-window-r2 0' to calculate R^2 , enabling R^2 values even below the default threshold of 0.2 to be reported. On average, this resulted in 142,249 R^2 values for the 0.05–0.275 MAF range, and in 241,154 R^2 values for the 0.275–0.5 MAF range. We then assigned the R^2 values to 1-kb bins according to the physical distance between the two focal SNPs, and plotted the mean R^2 for each bin from 1 to 100 kb. For the analysis of genome-wide LD decay with the full MAF range (0.05–0.5), we pooled the two MAF range specific PLINK outputs (one generated for the 0.05–0.275 and one for the 0.275–0.5 MAF range) before binning. Setting a MAF range of 0.05–0.5 right at the filtering step of the raw SNP matrix produced very similar results. To investigate more localized LD along chromosomes, we considered only R^2 values between SNPs > 2 kb but < 50 kb apart (a range between 2 kb and 30 kb, or considering pairwise R^2 values only produced similar results supporting identical conclusions). We determined the physical midpoints for all SNP pairs, binned the respective R^2 values in non-overlapping 200-kb windows along the genome, and calculated average R^2 for each window and population. Different window sizes (that is, 50 or 100 kb) yielded similar results supporting identical conclusions. To visualize localized differences in LD along the genome between the lake and GRA populations, we subtracted for each window the GRA R^2 value from its lake counterpart, yielding a metric referred to as 'Delta R^2 '. We further calculated the correlation of R^2 values between the lake and each stream population, using the above windows as data points. The magnitude of this correlation was evaluated against its empirical random distribution generated by permuting the R^2 data over the windows 10,000 times.

F_{ST} -based identification of selected regions. Scans for genomic regions exhibiting strong differentiation were performed for each lake–stream combination. (We decided to refer to particularly high differentiation values as 'extremes' rather than 'outliers', as the outlier terminology implies a distinct class of loci.) Consistent with refs 5,12, F_{ST} was calculated based on haplotype diversity. We considered only polymorphisms exhibiting a nucleotide coverage of at least $21\times$ in each population. To achieve adequate information to calculate genetic differentiation⁶², we further ignored SNPs with a MAF < 0.2 across the focal lake and stream population pool. If multiple SNPs derived from the same RAD locus, we selected only the single one yielding the highest F_{ST} value (selecting instead based on maximum MAF, or at random, had no material influence on the results). Applying these stringent filters, we obtained 55,476, 57,119 and 60,052 genome-wide F_{ST} values for the BOH–lake, NID–lake and GRA–lake comparisons. To obtain regions suited for a detailed characterization of signatures of selection, we chose the 25 autosomal SNPs displaying the highest F_{ST} values across the three F_{ST} data sets combined (that is, 172,647 F_{ST} estimates from 79,770 unique SNPs). To ensure that each of these differentiation extremes represented an independent genomic region, SNPs were ignored if they were closer than 200 kb to a SNP already accepted as extreme.

Haplotype-based identification of selected regions. Our F_{ST} -based search for evidence of positive selection was complemented with haplotype-based statistics proving particularly powerful to detect incomplete selective sweeps^{32,33}. However, they rely on relatively high marker resolution and robust sequence coverage in many individuals; requirements met by our study (see above and Supplementary Table 2). From the raw SNP matrix, we first excluded SNPs that were tri-allelic, had $> 40\%$ missing genotypes, or did not reach a MAF of 0.05. We further excluded individuals with $> 75\%$ missing diploid genotype calls after SNP-filtering. *fastPHASE* (ref. 63) was then used to reconstruct haplotypes and missing genotypes separately for each chromosome. We classified individuals according to their population (-u option) and increased the number of iterations of the EM algorithm to 50 (-C option; default is 25) and the number of sampled haplotypes to 100 (-H option; default is 20). *fastPHASE* output files were then imported into the R package *rehh* (ref. 64) to obtain the following haplotype-based statistics: EHH⁶⁵ (allele-specific 'Extended Haplotype Homozygosity'), EHHs⁶⁵ (population-specific weighted average of EHH across both alleles), iHH⁶⁶ ('integrated Haplotype Homozygosity'), iHS⁶⁶ ('integrated Haplotype Score') and Rsb³⁴ (the standardized ratio of integrated EHHs from two populations). iHS was calculated separately for each of the four populations using the 'scan_hh' and 'ihh2ihh' commands ('minmaf', the MAF threshold, was set to 0.05; '-freqbin' was set to 0, but setting this option to 0.05 or 0.1 resulted in qualitatively similar results

supporting identical conclusions). R_{sb} was calculated for each of the three possible lake–stream comparisons by applying default parameters ('ies2rsb' command). We obtained a total of 87,738 R_{sb} values (corresponding to an average marker distance of 4.8 kb), which were screened for extremes (that is, values below -4 or above 4 (refs 34,35)). Haplotype decay around F_{ST} and R_{sb} extremes was calculated and visualized using the 'calc_ehh' and the 'calc_ehhs' option at default.

Analyses specific to lateral plating. To screen the genome for loci influencing the lateral plate phenotype, we performed a BSA by assigning 24 completely plated stream individuals to one phenotypic group, and 24 low-plated stream individuals to another group (lateral plate phenotyping followed ref. 23 and is presented in Supplementary Table 2). This assignment considered all three stream populations but ignored the (mostly completely plated) lake fish, thus avoiding confounding signals of lake–stream divergence (that is, signals unrelated to plate phenotype). Based on 61,822 SNPs, we then carried out a genome-wide F_{ST} scan by treating the phenotypic groups as populations, but otherwise following all conventions described above for the population-based F_{ST} calculations.

To examine if *Eda* was recognized as a selected locus in a standard F_{ST} scan, we applied *BayeScan*⁴⁰ to the SNP data set from the GRA-lake comparison (60,052 markers), that is, the population pair with the strongest differentiation at *Eda*. *BayeScan* was run with default settings except that we used 300 as prior odds for neutrality—according to the software manual an appropriate value for this data set. However, a second analysis was performed with the default prior odds of 10, which is expected to produce more liberal results.

For the *Eda*-specific analyses of haplotype structure, we created three pools: a first pool with all completely plated stream individuals, a second pool with all low-plated stream individuals (both $N = 24$), and a third pool with all completely plated lake individuals ($N = 19$). We calculated and plotted EHS for each pool around the SNP exhibiting the highest F_{ST} value in the above bulk segregant genome scan (bp-position 12,832,658 on chromosome IV). Finally, we subtracted iHS values from the completely plated stream individuals from the corresponding values in the low-plated stream individuals ('Delta iHS' across chromosome IV ($N = 5,626$; average marker distance = 6 kb). Delta iHS was then averaged and plotted in non-overlapping 100-kb windows (different window sizes led to identical conclusions).

Identification and characterization of inversions. Our approach to detecting inversions was based on the expectation that the two inversion types (collinear and inverted), representing two isolated populations, differ in their magnitude of divergence from the reference genome. This should cause differential read alignment success across inverted genomic regions. Inversions should thus be revealed by a physically extended distortion of the relative RAD locus sequence coverage between two populations if these populations differ in the frequency of the inversion types (Supplementary Fig. 11). The same logic recently enabled the identification of evolutionary strata on the stickleback sex chromosome⁴². We therefore screened all 372,884 RAD loci for population-specific sequence coverage, excluding those with a total sequence coverage below 200 across all populations and those located in genomic regions unanchored to chromosomes, thus obtaining 290,170 informative loci. For each stream population, we calculated the RAD locus-specific stream to lake coverage ratio. Next, we divided the chromosomes in non-overlapping 20-kb windows (21,048 in total) and calculated the average coverage ratio among the RAD loci for each one of them (using the coverage variance among RAD loci within windows produced very similar results). The median number of RAD loci per window was 13. Finally, we looked for distortions in the coverage ratio extending over multiple adjacent windows, suggesting the presence of an inversion. We note that this analysis based on read coverage was limited to the detection of relatively large inversions exhibiting substantial sequence divergence.

To confirm that the above sequence coverage method reliably detects inversions, we used RAD loci near an expected inversion breakpoint in two of the three emerging candidate regions to design PCR primer pairs across the breakpoint boundaries. These primer pairs were expected to yield a PCR product only for the inversion type occurring in the streams. Ten to 13 individuals representing a given inversion type were subjected to long-range PCR and inspected for the presence or absence of amplification (further details are given in Supplementary Fig. 12).

Next, we examined allelic diversity and minor allele frequencies (MAFs) around the three detected inversions. For this, we screened each of the four population samples separately for polymorphisms with $> 50\%$ available genotype calls (singletons were omitted to exclude technical artifacts) and calculated haplotype diversity (that is, an analogue of heterozygosity ranging from 0 to 0.5) and the MAF at each SNP. RAD loci were allowed to contribute a single SNP only, keeping the one with the highest diversity when multiple SNPs occurred on the same locus (drawing a SNP at random or averaging diversity estimates of multiple SNPs per RAD locus yielded very similar results). Diversity was visualized using R's implementation of LOESS (locally weighted scatterplot smoothing; LOESS was used for all smoothing in this paper). The MAF frequency distribution within the inversions was plotted for the lake and for the stream population displaying the strongest inversion frequency shift from the lake. For this population, we also plotted the genome-wide MAF distribution.

To investigate LD patterns around the three inversions and to refine their physical boundaries, we calculated LD as the correlation among unphased SNP alleles using the R^2 statistic implemented in *mcl* (ref. 67). Only bi-allelic SNPs with $< 25\%$ missing data and individuals with $< 50\%$ missing *diploid* genotype calls were considered. When multiple SNPs were located on sister RAD loci, only a single randomly picked SNP was retained. For the calculation of LD, we applied different MAF filters, including a 0.15 MAF range centred on the MAF peak reflecting the relative frequency of the two variants at each inversion (see MAF analysis above). Patterns of LD around the inversions were visualized using the *LDheatmap* (ref. 68) R package for the stream population displaying the strongest inversion frequency shift from the lake (analysing the other stream populations yielded very similar estimates of the inversion breakpoint positions).

To construct haplotype genealogies for the inversions using individuals from the Lake Constance basin only, we first extracted the SNPs in each inversion, (SNPs closer than 20 kb to the inversion breakpoints identified in the LD analysis above were not considered). Next, we excluded SNPs with a MAF < 0.05 and with $> 25\%$ missing genotypes. Different MAF ranges (that is, 0.1–0.5 or 0.2–0.4) led to identical conclusions. Individuals with $> 75\%$ missing *diploid* genotypes after removing low-quality SNPs were excluded. When multiple SNPs per sister RAD loci passed the above filters, we only retained the one with the highest MAF (choosing a random SNP yielded similar results). For the largest inversion (located on ChrXXI), we randomly subsampled the resulting SNP panel to a total of 173 SNPs to reduce complexity. Haplotype reconstruction used *PHASE 2.1* (ref. 69), optimized by specifying the physical position of all polymorphisms and increasing the number of search iterations to 499. Five independent runs were performed with different seeds to confirm consistency among the results. Haplotype alignments were used to infer phylogenetic trees with *RAxML v.8.0.0* (ref. 70), using the GTRCAT model of sequence evolution with rate heterogeneity among sites. Based on sequence alignments and phylogenetic trees, we constructed and visualized haplotype genealogies with *Fitchi* (Matschiner, M.: *Fitchi: Haplotype genealogy graphs based on Fitch distances*. <http://www.evoinformatics.eu/fitchi>, 2015), using a minimal node size of two haplotypes for display ($-n$ option). To construct haplotype networks including individuals from across the stickleback's geographic range, we randomly selected 20 SNPs from the Lake Constance-specific haplotype genealogies, and inferred the genotypes at these SNPs for a total of 11 freshwater and 10 marine stickleback specimens¹¹ based on the ENSEMBL and the UCSC Stickleback Genome Browsers. The resulting SNPs (12, 13 and 14 for the ChrI, ChrXI and ChrXXI inversions) were used for haplotype network construction and visualization as described above.

The Lake Constance-specific haplotype networks allowed us to unambiguously infer *diploid* genotypes at all three inversions for our main study individuals. Of these individuals, 33 had already been RAD sequenced previously using the *SbfI* restriction enzyme⁶², allowing us to determine SNPs on *SbfI* RAD loci diagnostic for the two variants at each inversion. At these diagnostic SNPs, we then determined the *diploid* genotypes in 27 lake and 27 stream stickleback from the Lake Geneva basin²¹. For the stream individuals, *SbfI* RAD data were already available³⁸. For the Lake Geneva individuals, however, RAD sequence data were generated specifically for this study, following the protocol described in ref. 5. The SNP data from all individuals from the Lake Geneva basin were then used to search for the presence of inversion polymorphisms in this lake–stream system, to determine the frequencies of the inversion types in each population, and additionally to conduct an F_{ST} -based lake–stream genome scan.

To explore the short-term recombination rate at the inversions, we inspected genotype data from an F2 laboratory intercross⁴². This revealed that the two parental stickleback individuals used to initiate the cross (a male from Lake Constance and a female from a tributary stream of Lake Geneva) were fixed for different inversion types at the ChrI inversion (but not at the two other inversions). We therefore counted crossovers between SNPs across the ChrI inversion region in all 282 F2 individuals. As a negative control, we did the same around the ChrXI and ChrXXI inversions. To address the theoretical prediction that large inversions should maintain some genetic exchange due to double crossovers (gene conversion is considered less important)⁴⁷, we assigned stream individuals from the Lake Constance basin homozygous for one or the other inversion type at the ChrI inversion to separate groups ($N = 15$ and 20 for the stream and lake inversion type, defined according to Fig. 7c). These groups were then used to perform an F_{ST} -based differentiation scan. Additionally, using the same groups, we determined the number and location of loner SNPs specific to each inversion type, or shared between the types, within and around the ChrI inversion. Analogous analyses for the ChrXI and ChrXXI inversions were not possible because here individuals homozygous for the stream inversion type were too rare (Fig. 7c).

References

1. Arendt, J. & Reznick, D. Convergence and parallelism reconsidered: what have we learned about the genetics of adaptation? *Trends Ecol. Evol.* **23**, 26–32 (2008).
2. Conte, G. L., Arnegard, M. E., Peichel, C. L. & Schluter, D. The probability of genetic parallelism and convergence in natural populations. *Proc. Biol. Sci.* **279**, 5039–5047 (2012).
3. Martin, A. & Orgogozo, V. The Loci of repeated evolution: a catalog of genetic hotspots of phenotypic variation. *Evolution.* **67**, 1235–1250 (2013).

4. Tennessen, J. A. & Akey, J. M. Parallel adaptive divergence among geographically diverse human populations. *PLoS Genet.* **7**, e1002127 (2011).
5. 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).
6. Gagnaire, P.-A., Pavey, S. A., Normandeau, E. & Bernatchez, L. The genetic architecture of reproductive isolation during speciation-with-gene-flow in lake whitefish species pairs assessed by RAD sequencing. *Evolution* **67**, 2483–2497 (2013).
7. Foll, M., Gaggiotti, O. E., Daub, J. T., Vatsiou, A. & Excoffier, L. Widespread signals of convergent adaptation to high altitude in Asia and America. *Am. J. Hum. Genet.* **95**, 394–407 (2014).
8. Soria-Carrasco, V. *et al.* Stick insect genomes reveal natural selection's role in parallel speciation. *Science* **344**, 738–742 (2014).
9. Westram, A. M. *et al.* Do the same genes underlie parallel phenotypic divergence in different *Littorina saxatilis* populations? *Mol. Ecol.* **23**, 4603–4616 (2014).
10. Hohenlohe, P. A. *et al.* Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLoS Genet.* **6**, e1000862 (2010).
11. Jones, F. C. *et al.* The genomic basis of adaptive evolution in threespine sticklebacks. *Nature* **484**, 55–61 (2012).
12. 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).
13. Hardy, C. R. & Linder, H. P. Intraspecific variability and timing in ancestral ecology reconstruction: a test case from the Cape flora. *Syst. Biol.* **54**, 299–316 (2005).
14. Bierné, N., Gagnaire, P. A. & David, P. The geography of introgression in a patchy environment and the thorn in the side of ecological speciation. *Curr. Zool.* **59**, 72–86 (2013).
15. Barton, N. & Bengtsson, B. O. The barrier to genetic exchange between hybridizing populations. *Heredity* **57**, 357–376 (1986).
16. Bierné, N. The distinctive footprints of local hitchhiking in a varied environment and global hitchhiking in a subdivided population. *Evolution* **64**, 3254–3272 (2010).
17. Feder, J. L. & Nosil, P. The efficacy of divergence hitchhiking in generating genomic islands during ecological speciation. *Evolution* **64**, 1729–1747 (2010).
18. Endler, J. A. *Geographic variation, speciation, and clines* (Princeton University, 1977).
19. Coyne, J. A. & Orr, H. A. *Speciation* (Sinauer Associates, 2004).
20. Barrett, R. D. H. & Schluter, D. Adaptation from standing genetic variation. *Trends Ecol. Evol.* **23**, 38–44 (2008).
21. 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).
22. Lucek, K., Roy, D., Bezaul, E., Sivasundar, A. & Seehausen, O. Hybridization between distant lineages increases adaptive variation during a biological invasion: stickleback in Switzerland. *Mol. Ecol.* **19**, 3995–4011 (2010).
23. Moser, D., Roesti, M. & Berner, D. Repeated lake-stream divergence in stickleback life history within a Central European lake basin. *PLoS ONE* **7**, e50620 (2012).
24. Moser, D., Kueng, B. & Berner, D. Lake-stream divergence in stickleback life history: a plastic response to trophic niche differentiation? *Evol. Biol.* **42**, 328–338 (2015).
25. Bell, M. A. & Foster, S. A. *The evolutionary biology of the threespine stickleback* (Oxford University, 1994).
26. Reimchen, T. E. In *The evolutionary biology of the threespine stickleback* (eds Bell, M. A. & Foster, S. A.) 240–273 (Oxford University, 1994).
27. Keller, O. & Krauss, E. Die Hydrographie des Bodenseeraums in Vergangenheit und Gegenwart. *Ber. St. Gallen Nat.wiss. Ges.* **89**, 39–56 (2000).
28. Baird, N. A. *et al.* Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE* **3**, e3376 (2008).
29. Excoffier, L., Dupanloup, I., Huerta-Sanchez, E., Sousa, V. C. & Foll, M. Robust demographic inference from genomic and SNP data. *PLoS Genet.* **9**, e1003905 (2013).
30. Becquet, C. & Przeworski, M. Learning about modes of speciation by computational approaches. *Evolution* **63**, 1558–1566 (2009).
31. Cornuet, J.-M. *et al.* DIYABC v2.0: a software to make approximate Bayesian computation inferences about population history using single nucleotide polymorphism, DNA sequence and microsatellite data. *Bioinformatics* **30**, 1187–1189 (2014).
32. Sabeti, P. C. *et al.* Positive natural selection in the human lineage. *Science* **312**, 1614–1620 (2006).
33. Oleksyk, T. K., Smith, M. W. & O'Brien, S. J. Genome-wide scans for footprints of natural selection. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **365**, 185–205 (2010).
34. Tang, K., Thornton, K. R. & Stoneking, M. A new approach for using genome scans to detect recent positive selection in the human genome. *PLoS Biol.* **5**, e171 (2007).
35. Flori, L. *et al.* A quasi-exclusive European ancestry in the Senepol tropical cattle breed highlights the importance of the slick locus in tropical adaptation. *PLoS ONE* **7**, e36133 (2012).
36. Wu, C.-I. The genic view of the process of speciation. *J. Evol. Biol.* **14**, 851–865 (2001).
37. Colosimo, P. F. *et al.* Widespread parallel evolution in sticklebacks by repeated fixation of ectodysplasin alleles. *Science* **307**, 1928–1933 (2005).
38. Berner, D., Moser, D., Roesti, M., Buescher, H. & Salzburger, W. Genetic architecture of skeletal evolution in European lake and stream stickleback. *Evolution* **68**, 1792–1805 (2014).
39. O'Brien, N. M., Summers, B. R., Jones, F. C., Brady, S. D. & Kingsley, D. M. A recurrent regulatory change underlying altered expression and *Wnt* response of the stickleback armor plates gene *EDA*. *eLife* **4**, e05290 (2015).
40. Foll, M. & Gaggiotti, O. A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. *Genetics* **180**, 977–993 (2008).
41. Veit, G. *et al.* Collagen XXIII, novel ligand for Integrin $\alpha 2\beta 1$ in the epidermis. *J. Biol. Chem.* **286**, 27804–27813 (2011).
42. Roesti, M., Moser, D. & Berner, D. Recombination in the threespine stickleback genome—patterns and consequences. *Mol. Ecol.* **22**, 3014–3027 (2013).
43. Yeaman, S. Genomic rearrangements and the evolution of clusters of locally adaptive loci. *Proc. Nat. Acad. Sci. USA* **110**, 1743–1751 (2013).
44. Rieseberg, L. H. Chromosomal rearrangements and speciation. *Trends Ecol. Evol.* **16**, 351–358 (2001).
45. Kirkpatrick, M. & Barton, N. Chromosome inversions, local adaptation and speciation. *Genetics* **173**, 419–434 (2006).
46. Sturtevant, A. H. & Beadle, G. W. The relations of inversions in the X chromosome of *Drosophila melanogaster* to crossing over and disjunction. *Genetics* **21**, 554–604 (1936).
47. Navarro, A., Betran, E., Barbada, A. & Ruiz, A. Recombination and gene flux caused by gene conversion and crossing over in inversion heterokaryotypes. *Genetics* **146**, 695–709 (1997).
48. Guerrero, R. F., Rousset, F. & Kirkpatrick, M. Coalescent patterns for chromosomal inversions in divergent populations. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **367**, 430–438 (2012).
49. Schaeffer, S. W. & Anderson, W. W. Mechanisms of genetic exchange with the chromosomal inversions of *Drosophila pseudoobscura*. *Genetics* **171**, 1729–1739 (2005).
50. Stump, A. D. *et al.* Genetic exchange in 2La inversion heterokaryotypes of *Anopheles gambiae*. *Insect. Mol. Biol.* **16**, 703–709 (2007).
51. Pearce, D. E., Miller, M. R., Abadia-Cardoso, A. & Garza, J. C. Rapid parallel evolution of standing variation in a single, complex, genomic region is associated with life history in steelhead/rainbow trout. *Proc. R. Soc. B* **281**, 20140012 (2014).
52. Flaxman, S. M., Wacholder, A. C., Feder, J. L. & Nosil, P. Theoretical models of the influence of genomic architecture on the dynamics of speciation. *Mol. Ecol.* **23**, 4074–4088 (2014).
53. Feder, J. L., Egan, S. P. & Nosil, P. The genomics of speciation-with-gene-flow. *Trends Genet.* **28**, 342–350 (2012).
54. Cruickshank, T. E. & Hahn, M. W. Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Mol. Ecol.* **23**, 3133–3157 (2014).
55. Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
56. Nevado, B., Ramos-Onsins, S. E. & Perez-Enciso, M. Resequencing studies of nonmodel organisms using closely related reference genomes: optimal experimental designs and bioinformatics approaches for population genomics. *Mol. Ecol.* **23**, 1764–1779 (2014).
57. R Core Team. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria, 2013).
58. Schliep, K. P. Phangorn: Phylogenetic analysis in R. *Bioinformatics* **27**, 592–593 (2011).
59. Posada, D. jModelTest: Phylogenetic model averaging. *Mol. Biol. Evol.* **25**, 1253–1256 (2008).
60. Paradis, E., Claude, J. & Strimmer, K. Analyses of phylogenetics and evolution in R language. *Bioinformatics* **20**, 289–290 (2004).
61. Purcell, S. *et al.* PLINK: a toolset for whole-genome association and population-based linkage analysis. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
62. Roesti, M., Salzburger, W. & Berner, D. Uninformative polymorphisms bias genome scans for signatures of selection. *BMC Evol. Biol.* **12**, 94 (2012).
63. Scheet, P. & Stephens, M. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am. J. Hum. Genet.* **78**, 629–644 (2006).
64. Gautier, M. & Vitalis, R. rehh: An R package to detect footprints of selection in genome-wide SNP data from haplotype structure. *Bioinformatics* **28**, 1176–1177 (2012).

65. Sabeti, P. C. *et al.* Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419**, 832–837 (2002).
66. Voight, B. F., Kudaravalli, S., Wen, X. & Pritchard, J. K. A map of recent positive selection in the human genome. *PLoS Biol.* **4**, e72 (2006).
67. Zaykin, D. V., Pudovkin, A. & Weir, B. S. Correlation-based inference for linkage disequilibrium with multiple alleles. *Genetics* **180**, 533–545 (2008).
68. Shin, J.-H., Blay, S., McNeney, B. & Graham, J. LDheatmap: an R function for graphical display of pairwise linkage disequilibrium between single nucleotide frequencies. *J. Stat. Softw.* **16**, 1–8 (2006).
69. Stephens, M., Smith, N. J. & Donnelly, P. A new statistical method for haplotype reconstruction from population data. *Am. J. Hum. Genet.* **68**, 978–989 (2001).
70. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*. **30**, 1312–1313 (2014).

Acknowledgements

We thank Walter Salzburger (financial support and infrastructure); Brigitte Aeschbach and Nicolas Boileau (wet lab support); Christian Beisel and Ina Nissen (Illumina sequencing at the Quantitative Genomics Facility, D-BSSE, ETH Zurich); Steve Lianoglou, Martin Morgan and Hervé Pagès (suggestions to speed up the genotyping pipeline); Laurent Excoffier (demographic analysis); Matthew Stephens (haplotype reconstruction); Michael Matschiner (phylogenetic inference and haplotype networks); Mathieu Gautier (EHH-based analyses). M.R. was supported financially by a Swiss National Science Foundation (SNSF) Sinergia grant (CRSII3_136293) awarded to Walter Salzburger, Marcelo Sanchez and Heinz Furrer. D.B. was supported by the SNSF (Ambizione grant PZ00P3_126391/1 and project grant 31003A_146208/1), the University of Basel and the Freiwillige Akademische Gesellschaft Basel.

Author contributions

M.R. designed the study, conducted field sampling, did wet lab work, analysed the data, interpreted and visualized results, and contributed to paper writing. B.K. did wet lab work and contributed to data analysis. D.M. conducted field sampling and wet lab work. D.B. designed the study, conducted field sampling, developed the genotyping pipeline, analysed data, interpreted results and wrote the paper.

Additional information

Accession codes: RAD-seq data generated for this study have been deposited in the NCBI Sequence Read Archive under BioProject number PRJNA273792 (Nsi1 data from the Lake Constance region) and PRJNA284945 (Sbf1 data from the Lake Geneva region).

Supplementary Information accompanies this paper at <http://www.nature.com/naturecommunications>

Competing financial interests: The authors declare no competing financial interests.

Reprints and permission information is available online at <http://npg.nature.com/reprintsandpermissions/>

How to cite this article: Roesti, M. *et al.* The genomics of ecological vicariance in threespine stickleback fish. *Nat. Commun.* **6**:8767 doi: 10.1038/ncomms9767 (2015).



This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>

Erratum: The genomics of ecological vicariance in threespine stickleback fish

Marius Roesti, Benjamin Kueng, Dario Moser & Daniel Berner

Nature Communications 6:8767 doi: 10.1038/ncomms9767 (2015); Published 10 Nov 2015; Updated 15 Dec 2015

The recombination rate around the chromosomal inversion in Fig. 6f of this Article was inadvertently omitted during the production process. The correct version of Fig. 6f appears below.

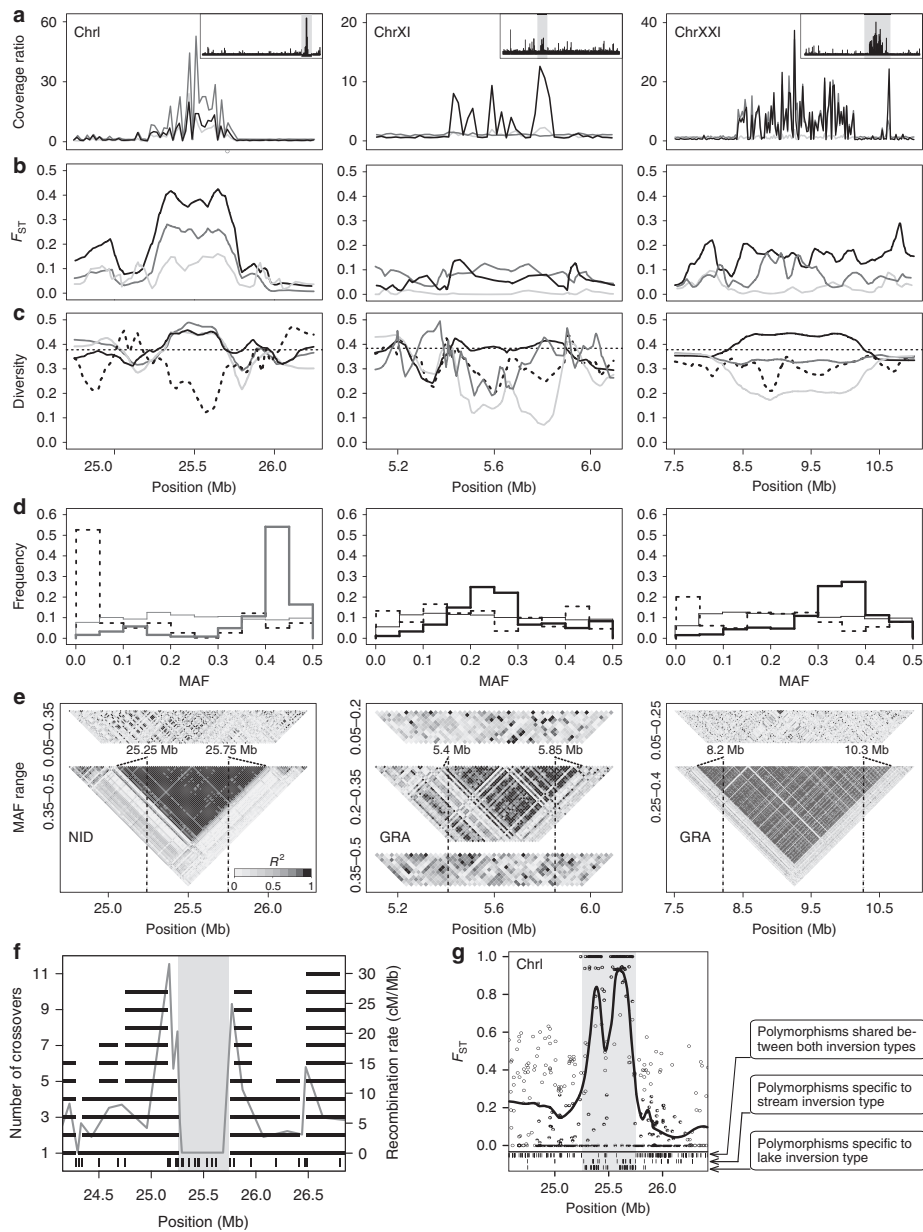



Figure 6.

 This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>

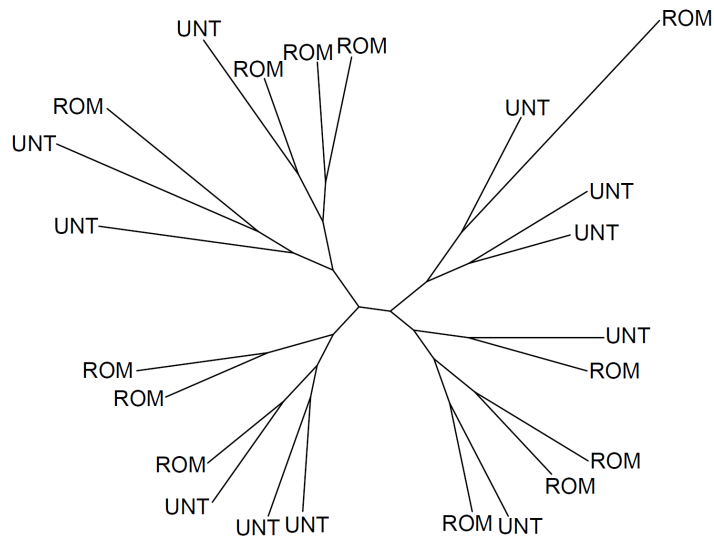
Chapter 6

The genomics of ecological vicariance in threespine
stickleback fish

Supplementary material

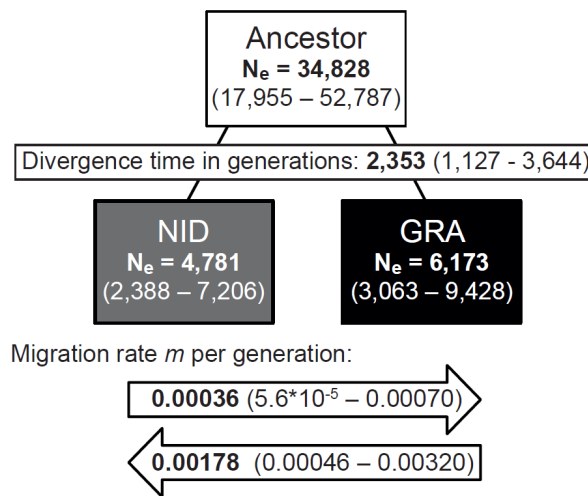
SUPPLEMENTARY FIGURES

Supplementary Figure 1



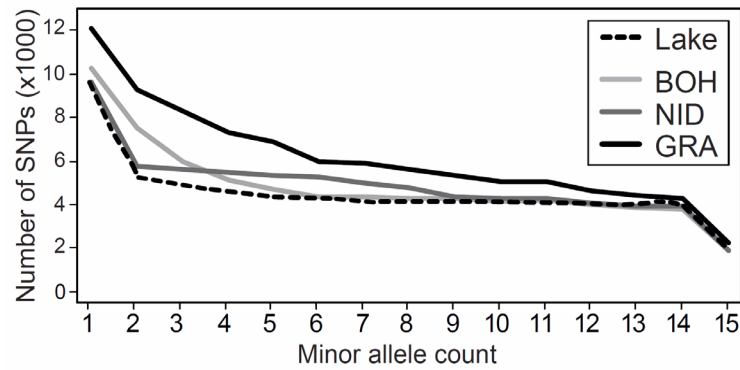
Supplementary Figure 1. Phylogenetic analysis restricted to the two stickleback samples from Lake Constance. The unrooted maximum likelihood tree is based on 55,561 genome-wide SNPs in fish sampled from two lake sites approximately 20 km apart. The sites are Romanshorn (ROM), Switzerland, western lake shore, and Unteruhldingen (UNT), Germany, eastern shore (for geographic details see ref. 1). Consistent with a genome-wide median F_{ST} of zero between ROM and UNT, the phylogeny reveals the absence of genetic structure between the two sites, indicating that Lake Constance is inhabited by a single panmictic stickleback population. The same conclusion was drawn earlier based on microsatellite markers and stickleback samples from four different lake sites¹.

Supplementary Figure 2



Supplementary Figure 2. Demographic analysis based on a reduced model including the GRA and NID stream populations only. Plotting conventions are as in the full model including all study populations (Fig. 1b). The GRA and NID populations are the genetically most variable of our study populations (see main text). In the reduced model, the split between GRA and NID from a common ancestor is estimated to have occurred more recently compared to the full model, although the confidence intervals overlap widely between the models. A potential reason for the deeper splitting time in the full model is upward bias due to extensive genome-wide selective sweeps experienced by the lake population. We thus consider the splitting time estimate from the reduced model a better approximation of the true time since stickleback colonized the Lake Constance basin. However, both models support qualitatively similar conclusions about the colonization history of the Lake Constance basin.

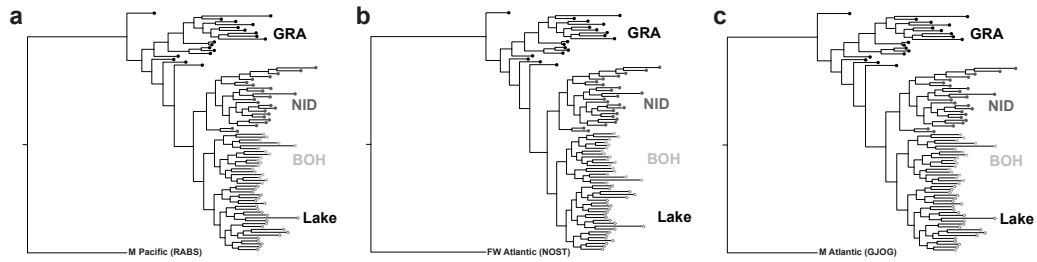
Supplementary Figure 3



Supplementary Figure 3. Observed site frequency spectra (SFS) of the four study populations.

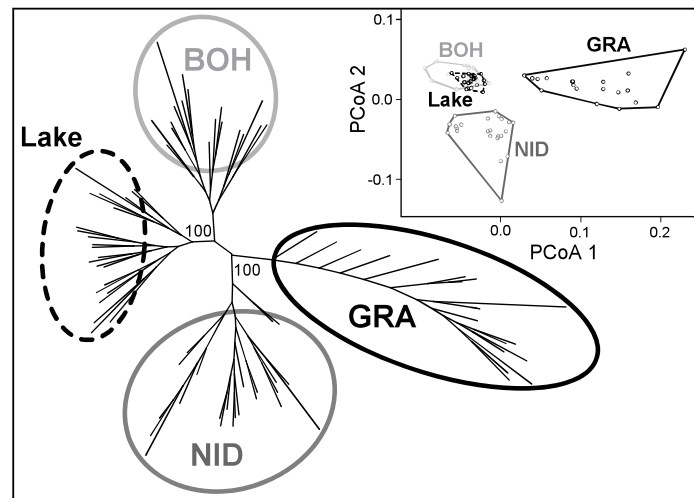
The X-axis indicates the occurrence of the minor allele among 30 randomly sampled nucleotides at a given genome position (the minor allele *frequency* (MAF) would thus be obtained by dividing the counts by 30). The Y-axis gives the number of sites falling into each minor allele count class in each population. Like the joint SFS used for demographic inference, these population-specific SFS are based on 14.8 million nucleotide positions, although for the ease of presentation, only the polymorphic sites (i.e., minor allele count > 0) are shown. Note the low number of polymorphisms across most minor allele count classes in the lake population relative to the stream populations (especially GRA). Accordingly, the lake population exhibited the highest proportion of monomorphic sites (minor allele count = 0); in millions, lake: 14.770; BOH: 14.765; NID: 14.764; GRA: 14.745.

Supplementary Figure 4



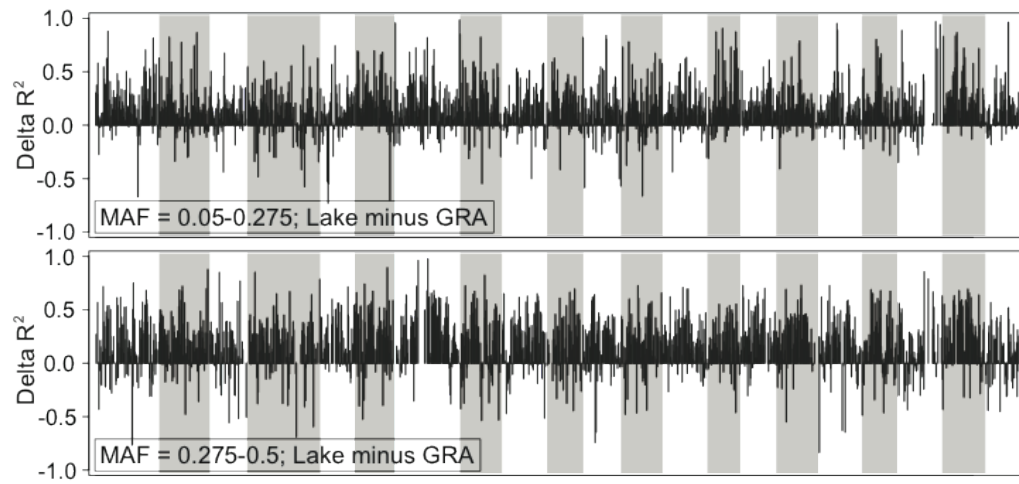
Supplementary Figure 4. Phylogenetic relationships among the Lake Constance study populations rooted using different outgroups. To verify the robustness of the rooted ML phylogeny by using the reference genome individual (a freshwater individual from the Pacific, see Fig. 1c) as an outgroup, we generated additional trees using several other outgroups, including **(a)** a marine Pacific (sampling population: 'Rabbit Slough', Alaska), **(b)** an freshwater Atlantic (sampling population: 'Norway Stream', Norway), and **(c)** a marine Atlantic (sampling population: 'Gjögur', Iceland) stickleback individual. Genotypes for these individuals were retrieved from the 'Stickleback Genome Browser' (<http://sticklebrowser.stanford.edu/>)². These analyses consistently resulted in very similar tree topologies supporting identical conclusions.

Supplementary Figure 5



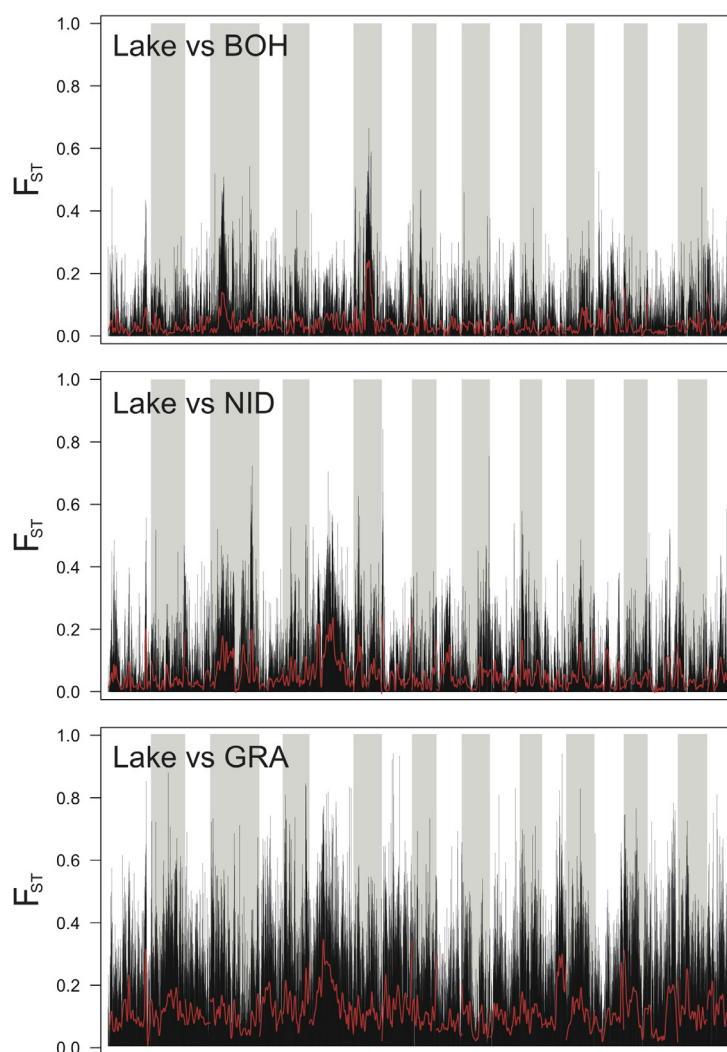
Supplementary Figure 5. Phylogenetic and principal coordinate analysis of the four stickleback populations from the Lake Constance basin. The unrooted maximum likelihood tree (based on 51,188 SNPs; bootstrap support in percent is given for the key nodes) reveals reciprocal monophyly of the four populations. Both the tree and the principal coordinate ordination (insert) further show the close relatedness of the lake and the BOH population, and that genetic diversity increases from the lake population to the BOH, NID, and GRA stream populations.

Supplementary Figure 6



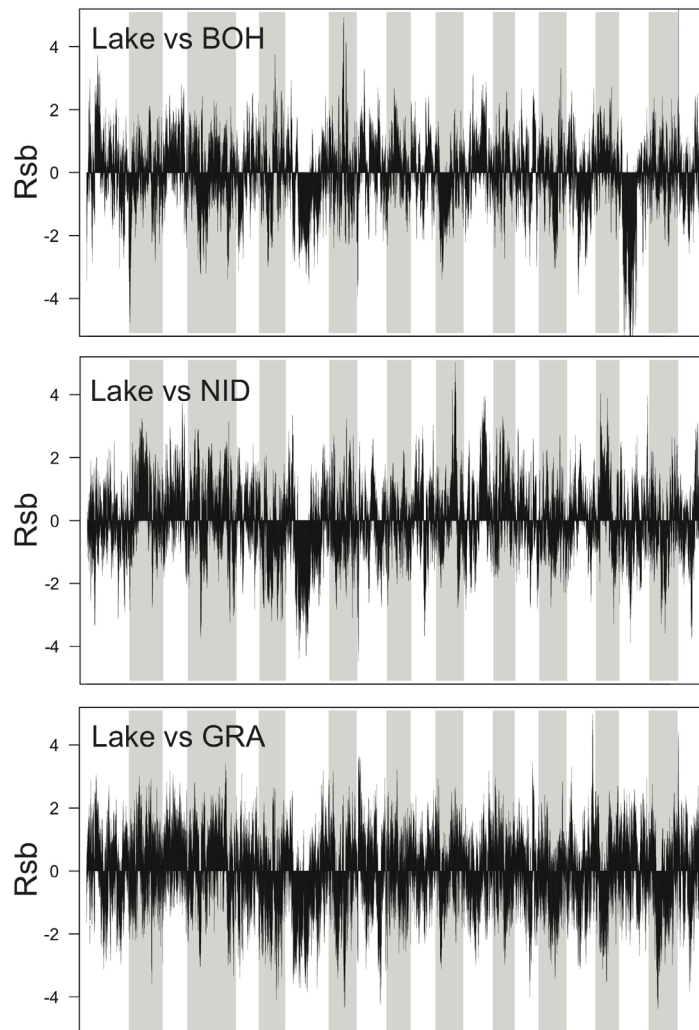
Supplementary Figure 6. Influence of using SNPs from different MAF classes on the difference in LD between the lake and the GRA population. Shown is Delta R² (see Fig. 2b) based on low-MAF (top) and high-MAF (bottom) SNPs. The MAF classes are separated using the same thresholds as used in the insert of Fig. 2a. Irrespective of the MAF class, LD is higher in the lake than in GRA along most of the genome.

Supplementary Figure 7



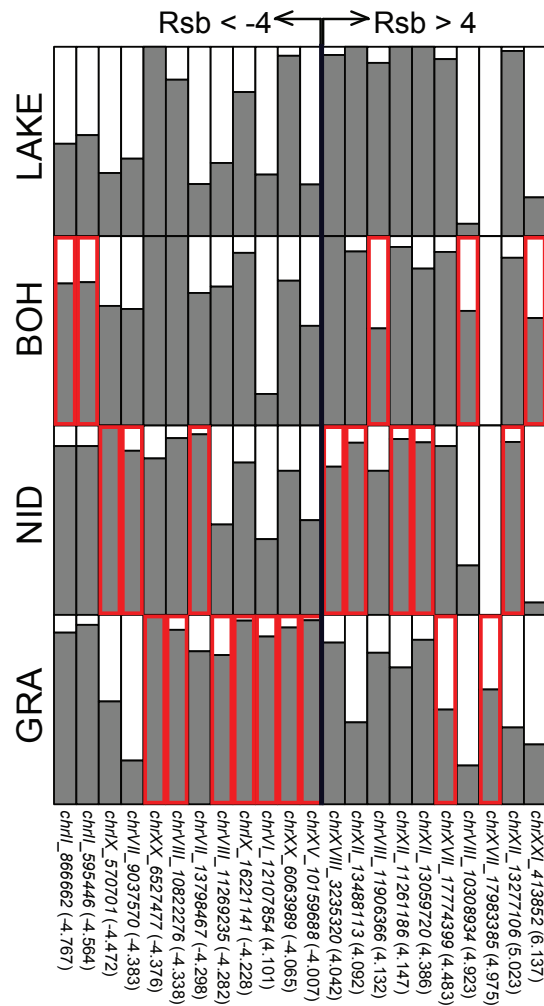
Supplementary Figure 7. Genome-wide divergence (F_{ST}) for all lake-stream comparisons. The black vertical lines represent the raw F_{ST} values, the red profiles show these values smoothed by LOESS, and the background shading separates the 21 chromosomes. Note the increase in baseline differentiation from BOH (median F_{ST} = 0.005; 55,476 SNPs) to NID (0.013; 57,119 SNPs) and GRA (0.061; 60,052 SNPs).

Supplementary Figure 8



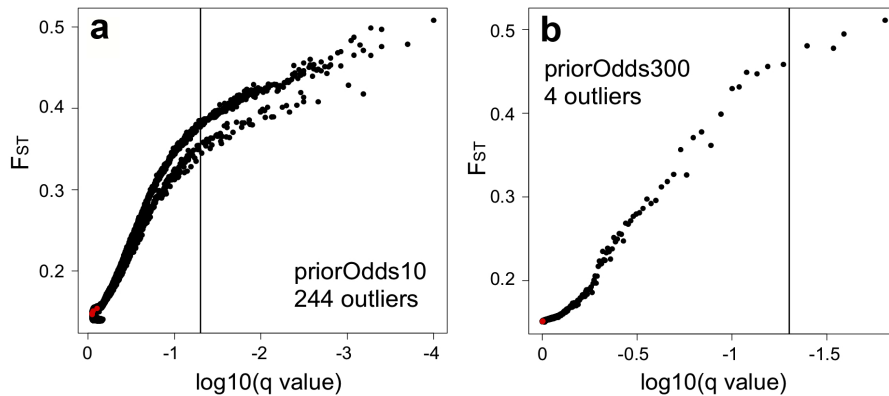
Supplementary Figure 8. Difference in haplotype decay around genome-wide SNPs, as captured by R_{sb} , for each lake-stream population pairing. The background shading separates the 21 chromosomes. A total of 87,738 SNPs were used in all lake-stream comparisons.

Supplementary Figure 9



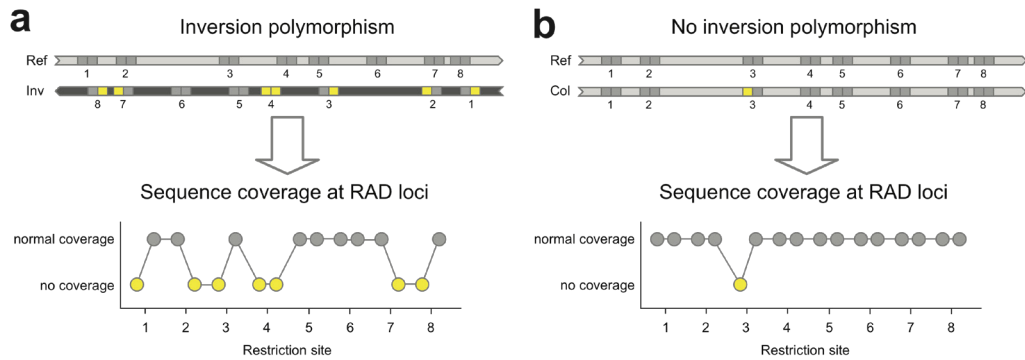
Supplementary Figure 9. Allele frequencies within each population at the top 22 lake-stream Rsb extremes. At each Rsb extreme (columns), the stream population producing an absolute Rsb > 4 in comparison to the lake is framed in red. On the bottom, the genomic position and the highest Rsb value observed across all lake-stream comparisons are given for each Rsb extreme. *Negative* Rsb extremes generally display relatively balanced polymorphism in the lake, but strong bias toward a specific allele in the stream(s), hence suggesting stream-specific selective sweeps. By contrast, *positive* Rsb extremes tend to exhibit relatively balanced polymorphism in the streams but are near fixation for a specific allele in the lake, thus indicating lake-specific selective sweeps.

Supplementary Figure 10



Supplementary Figure 10. BayeScan divergence outlier analysis in the Lake Constance and GRA stream population pair. Analysis to explore if markers near the Ectodysplasin (*Eda*) gene, known to be under divergent selection between these populations, are recognized as selection outliers by a popular outlier detection program not requiring a reference genome. The analysis used 60,052 SNPs, and was run both with default settings **(a)**, and with the prior odds for neutrality increased to 300 **(b)** (default is 10). According to the software manual, the latter setting should be more appropriate for our large marker data set, while the default is perhaps too liberal. The graphics display the results of these two outlier scans, with the five markers near *Eda* exhibiting the highest F_{ST} in our differentiation scan printed in red (see top panel in Fig. 5c; positions on ChrIV: 12,815,791; 12,818,350; 12,818,237; 12,820,744; 12,822,878). SNPs on the right of the vertical line (244 and 4 in the two scans) qualify as differentiation outliers at a false discovery rate of 0.05. None of the markers near *Eda* are identified as outliers by BayeScan.

Supplementary Figure 11



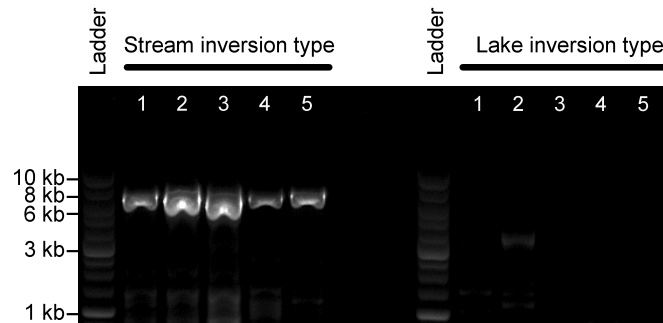
Supplementary Figure 11. Strategy for the detection of inversion polymorphisms using RAD locus coverage. **(a)** An individual harboring the inverted inversion type ('Inv'; dark gray background shading) relative to the reference sequence ('Ref'; light gray background) (for simplicity, individuals are haploid in this figure). The small squares represent the two RAD loci flanking restriction enzyme cutting sites to either side (sister RAD loci). If the 'Inv' inversion type shows substantial divergence from the reference, individuals carrying this type will lack sequence coverage at many RAD loci when aligned to the reference (RAD loci too strongly differentiated to align to the reference are shown as yellow squares). The bottom panel shows the resulting pattern of sequence coverage across RAD loci for this inversion type. **(b)** An individual carrying the inversion type collinear ('Col') to the reference (top), and the resulting sequence coverage along this chromosomal segment (bottom).

If the different inversion types segregate at different frequencies within two populations, mean sequence coverage across chromosome windows within the inversion will be biased toward the population in which the 'Inv' type is less common, relative to chromosome segments outside the inversion. An analogous signature emerges when comparing the variance in sequence coverage across chromosome windows within and outside inversions between populations. Both signals, i.e., bias in the ratio of mean sequence coverage and coverage variance between populations along the genome, were exploited in our study and both consistently detected the three inversions, although only the former is presented (Fig. 6a). (Note that distortions in mean coverage and coverage variance along chromosomes can also be used to detect inversions in a single population, although the comparison of populations provides additional information on shifts in inversion frequencies.)

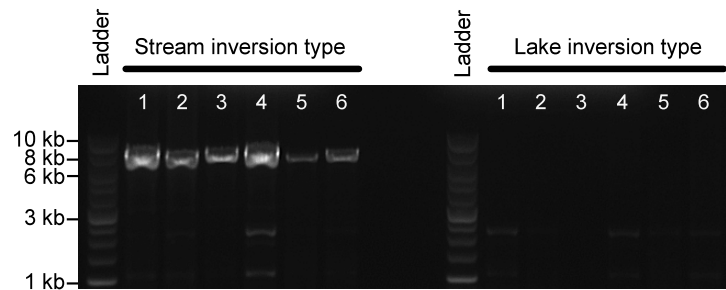
The prerequisites for the above inversion detection approaches are that the inverted and collinear segments display substantial sequence divergence (recent inversions cannot be detected), and that the density of restriction sites is high enough to allow calculating the bias in the ratio of mean sequence coverage or the coverage variance between populations in relatively small chromosome windows while still integrating coverage data from a reasonably large number of RAD loci (a low-frequency restriction enzyme digest will allow detecting large inversions only). Moreover, comparing coverage statistics between populations will detect inversion only when these populations have diverged sufficiently in the frequency of the inversion types.

Supplementary Figure 12

ChrI inversion



ChrXI inversion

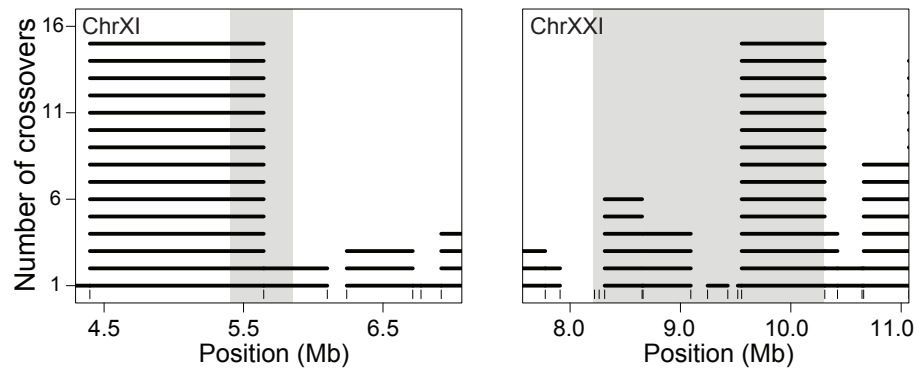


Supplementary Figure 12. Confirming inversions by inversion type-specific PCR across expected breakpoints. For the putative ChrI and ChrXI inversions detected based on RAD sequence coverage, we used RAD loci flanking one inversion breakpoint to design PCR primer pairs expected to yield a PCR product for the inversion type specific to the streams, but no product for the inversion type fixed in the lake (see Fig. 7c). The underlying RAD loci were required to display robust alignment to the reference genome in *all* populations, thus ensuring that any absence of PCR amplification was due to the physical relocation of a primer site, and not to the degeneration of a primer site. For the ChrI inversion, we assessed 13 individuals homozygous for the stream type, of which nine (70%) amplified successfully, and ten individuals homozygous for the lake type, of which none amplified (five individuals of each group are shown on the gel image). For the ChrXI inversion, we assessed five individuals homozygous for the stream type and seven heterozygous

individuals, all of which amplified successfully. By contrast, none of the ten individuals homozygous for the lake type amplified (six individuals of each group are visualized; the individuals 5 and 6 in the stream inversion group are heterozygous). These analyses thus confirm that the candidate regions are truly inversions. Note that the ChrI (and also the ChrXXI) inversion has been confirmed independently through PCR, using different primer pairs than in the present study². The ChrXI inversion, however, has not previously been verified by PCR.

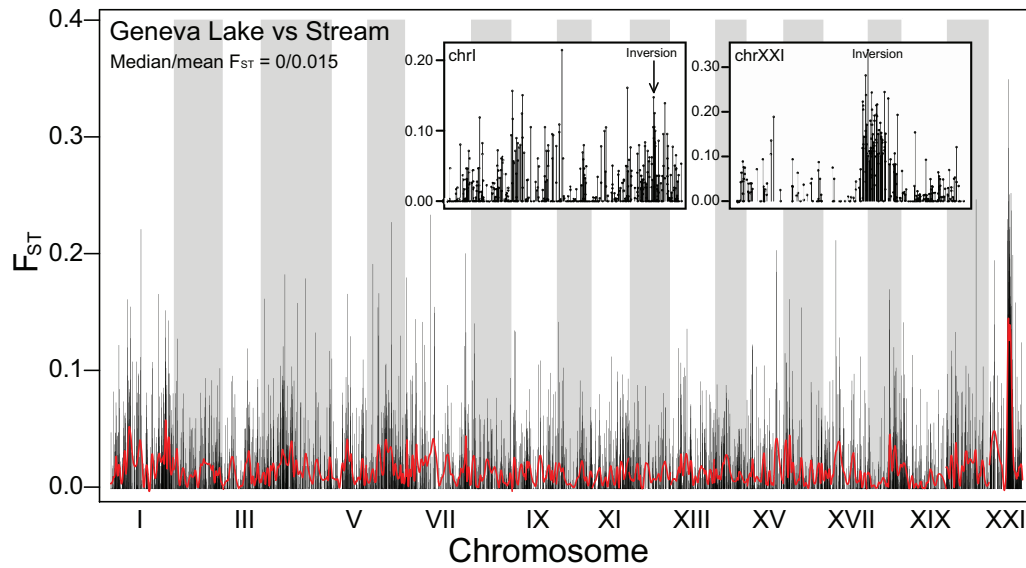
As representatives of both inversion types, our PCRs considered primarily individuals from the stream populations in the Lake Constance basin (these populations are polymorphic for the inversions; Fig. 7). A few individuals from the Lake Geneva basin (Fig. 7c), however, were included in all reactions, which showed that geographic origin did not influence amplification success. The primer combinations used for this analysis were 5'- GCTGGTCAATATGTCCACTC-'3 (forward) and 5'- GTTACAATATGCCAATTACATGTC-'3 (reverse) for ChrI (approximate expected product size: 6.2 kb), and 5'-GGAGAAGCCTCAACCTATACG-'3 (forward) and 5'-GGTGAGCAACTTGAACCAAG-'3 (reverse) for ChrXI (6.8 kb). Long-range PCRs were performed with 37 cycles using Phusion High-Fidelity PCR chemistry (New England BioLabs), following the manufacturer's protocol.

Supplementary Figure 13



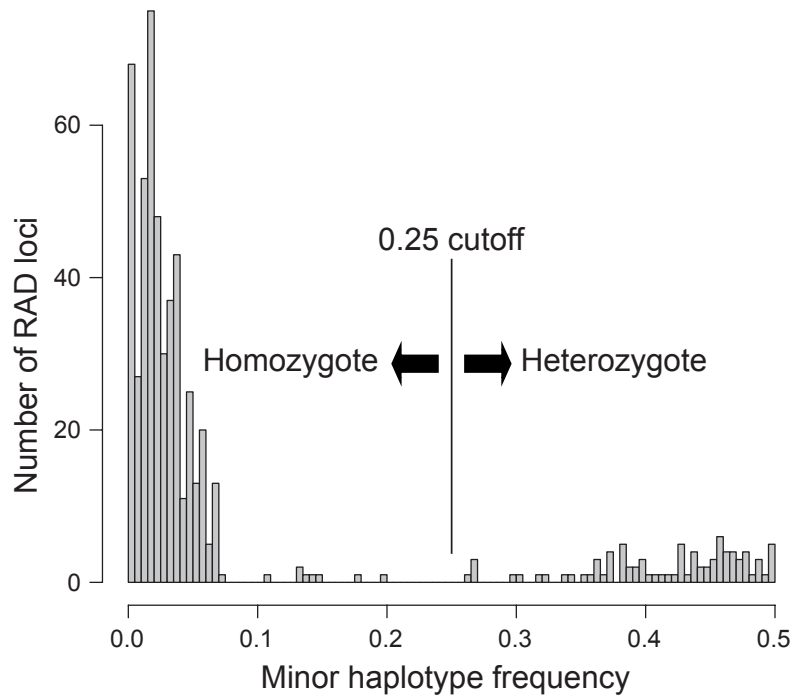
Supplementary Figure 13. Recombination rate around the ChrXI and ChrXXI inversions in a laboratory cross population. Plotting conventions are as in Fig. 6f. For the ChrXI and ChrXXI inversion, the cross population underlying the recombination analysis reported in Fig. 6f is monomorphic. We here show that, as expected, recombination in these regions is *not* suppressed, thus providing a negative control for the analysis presented in Fig. 6f.

Supplementary Figure 14



Supplementary Figure 14. Genetic differentiation (F_{ST}) between the Lake Geneva population, and a stream population from a tributary of this lake. The black vertical lines represent the raw lake-stream F_{ST} values, the red profiles show these values smoothed by LOESS, and the background shading separates the 21 chromosomes. The genome region displaying the strongest differentiation is located on ChrXXI and coincides with the large inversion on that chromosome (right insert; average F_{ST} across this inversion: 0.160). Relative to the low genome-wide baseline differentiation (given in top-left corner), the ChrI inversion also exhibits strong lake-stream divergence (left insert; average F_{ST} across this inversion: 0.084).

Supplementary Figure 15



Supplementary Figure 15. Determining an appropriate threshold for distinguishing homozygote and heterozygote RAD loci during consensus genotyping. To identify this threshold, we determined the frequency of all haplotypes occurring at 250 haphazardly chosen RAD loci in each of three individuals displaying low, medium, and high raw Illumina sequence coverage. Among these 750 total RAD loci, we discarded those in which the two most frequent haplotypes together failed to account for > 70% of all haplotypes and/or to reach a sum of 15 (see Methods). Across the remaining 562 RAD loci, we then calculated the minor haplotype frequency, defined as the count of the second most frequent haplotype divided by the sum of the two most frequent haplotypes. The distribution of this statistic indicated that a cutoff around 0.25 effectively separated truly heterozygous RAD loci from those appearing variable because of a technical artifact.

SUPPLEMENTARY TABLES

Supplementary Table 1. Genetic diversity within each of the four study populations. Diversity is calculated based exclusively on 'loner SNPs' (i.e., SNP occurring alone on their RAD locus, see Methods). The first two data columns indicate the number and corresponding proportion of the total loner SNPs (N = 62,332) actually being polymorphic within each population (in parentheses the proportions are scaled such that the lake is 100%). This proportion is lowest in the lake population. Analogously, the third and fourth data columns report the number and proportion of the total tri-allelic loner SNPs (N = 368) actually being tri-allelic within each population (in parentheses the proportions are scaled as above). This latter diversity index is again lowest in the lake population.

Population	Number of loner SNPs polymorphic in focal population	Proportion of loner SNPs polymorphic in focal population	Number of tri-allelic loner SNPs polymorphic in focal population	Proportion of tri-allelic loner SNPs polymorphic in focal population
Lake	44,070	0.707 (100.0%)	103	0.280 (100.0%)
BOH	45,838	0.735 (104.0%)	97	0.264 (94.2%)
NID	46,632	0.748 (105.8%)	126	0.342 (122.3%)
GRA	56,280	0.903 (127.7%)	188	0.511 (182.5%)

Supplementary Table 2. Descriptive statistics for the study individuals.

Illumina library	Barcode	Specimen ID	Population	Sex	Lateral plate phenotype (l = low; p = partial; f = full)	Number of raw Illumina reads	Alignment success	Number of RAD loci yielding a consensus genotype	Mean sequence coverage across the RAD loci yielding a consensus genotype
Ga_lib_21	CGATA	Ga1085	GRA	f	p	16'123'901	0.83	287'145	43.1
Ga_lib_21	CGGCG	Ga1170	BOH	f	p	7'024'520	0.81	276'801	18.0
Ga_lib_21	CTAGG	Ga1172	BOH	f	N.A.	9'782'486	0.82	283'283	25.4
Ga_lib_21	CTGAA	Ga1426	GRA	m	p	9'224'793	0.82	284'721	24.1
Ga_lib_21	GAAGC	Ga1082	GRA	m	l	14'974'371	0.82	287'020	39.2
Ga_lib_21	GAGAT	Ga1152	BOH	f	f	4'456'312	0.82	259'253	12.1
Ga_lib_21	GCATT	Ga1173	BOH	f	N.A.	20'910'550	0.81	285'786	53.5
Ga_lib_21	GCGCC	Ga1119	UNT	m	f	27'416'119	0.78	283'349	67.5
Ga_lib_22	GGAAG	Ga1156	BOH	m	f	16'926'683	0.70	286'303	36.6
Ga_lib_22	GTACA	Ga1081	GRA	f	l	27'433'412	0.62	287'802	53.3
Ga_lib_22	TAATG	Ga1078	GRA	m	f	5'127'972	0.74	257'782	12.3
Ga_lib_22	TAGCA	Ga1110	UNT	f	p	5'164'073	0.70	266'488	12.1
Ga_lib_22	TCAGA	Ga1103	UNT	m	f	39'406'778	0.61	285'140	74.1
Ga_lib_22	TCGAG	Ga1106	UNT	f	f	2'848'858	0.71	221'902	7.6
Ga_lib_22	TGACC	Ga0312	NID	f	p	23'161'376	0.58	288'084	42.3
Ga_lib_22	TGGTT	Ga1120	UNT	m	f	20'280'307	0.54	279'849	35.0
Ga_lib_23	CGCGC	Ga0070	NID	f	l	36'334'680	0.63	274'580	73.6
Ga_lib_23	CGTAT	Ga0314	NID	m	l	23'496'158	0.64	288'445	47.9
Ga_lib_23	CTCTT	Ga1168	BOH	m	f	26'380'282	0.63	290'415	51.8
Ga_lib_23	CTTCC	Ga0087	ROM	f	f	38'114'299	0.61	276'609	73.8
Ga_lib_23	GACTA	Ga0337	NID	m	p	27'011'624	0.65	274'782	56.3
Ga_lib_23	GATCG	Ga1154	BOH	m	f	11'986'012	0.66	282'398	25.3
Ga_lib_23	GCCGG	Ga1114	UNT	m	p	11'363'641	0.67	281'197	24.2
Ga_lib_23	GCTAA	Ga0090	ROM	m	f	24'690'056	0.65	274'506	51.7
Ga_lib_24	GGCCT	Ga1155	BOH	m	l	6'711'057	0.73	252'208	16.4
Ga_lib_24	GGTTC	Ga0144	NID	f	p	30'968'137	0.67	278'899	66.1
Ga_lib_24	GTCAC	Ga1100	UNT	f	f	19'294'137	0.72	281'433	44.5
Ga_lib_24	GTTGT	Ga1427	GRA	m	l	6'193'909	0.71	262'212	14.0
Ga_lib_24	TATAC	Ga1160	BOH	m	f	25'348'324	0.70	286'704	54.5

Ga_lib_24	TCCTC	Ga0159	NID	f	f	35'699'372	0.69	279'026	79.2
Ga_lib_24	TGCAA	Ga1166	BOH	m	p	18'887'595	0.71	284'320	42.4
Ga_lib_24	TGTGG	Ga1076	GRA	m	p	6'508'538	0.69	260'893	14.7
Ga_lib_25	CGATA	Ga1104	UNT	f	f	19'778'122	0.68	284'455	42.4
Ga_lib_25	CGGCG	Ga0094	ROM	m	f	23'777'863	0.63	276'723	48.3
Ga_lib_25	CTAGG	Ga0074	NID	m	l	36'507'287	0.64	280'769	75.7
Ga_lib_25	CTGAA	Ga0228	NID	m	p	39'198'188	0.63	279'035	80.2
Ga_lib_25	GAAGC	Ga1074	GRA	m	p	14'214'345	0.68	283'572	30.2
Ga_lib_25	GAGAT	Ga1158	BOH	m	l	10'332'558	0.68	279'500	22.1
Ga_lib_25	GCATT	Ga1098	UNT	m	f	20'686'598	0.67	286'410	43.8
Ga_lib_25	GCGCC	Ga0075	NID	m	f	33'589'851	0.62	275'988	68.4
Ga_lib_26	GTACA	Ga1080	GRA	f	l	14'841'946	0.66	282'456	31.1
Ga_lib_26	CGATA	Ga0120	NID	m	p	49'283'123	0.64	277'554	102.7
Ga_lib_26	GAGAT	Ga1167	BOH	m	f	20'724'513	0.66	286'797	42.1
Ga_lib_26	TCGAG	Ga0078	NID	f	f	53'009	0.54	1'097	2.1
Ga_lib_26	ACACG	Ga0146	ROM	f	f	29'091'722	0.64	276'402	60.6
Ga_lib_26	AGAGT	Ga1107	UNT	f	f	32'492'310	0.66	285'290	67.5
Ga_lib_26	CATGA	Ga1083	GRA	f	p	8'164'759	0.72	278'025	18.4
Ga_lib_26	ATGCT	Ga0201	ROM	f	f	37'919'120	0.63	279'504	76.2
Ga_lib_27	TAATG	Ga0114	ROM	m	p	24'970'277	0.60	289'170	45.8
Ga_lib_27	CGATA	Ga1084	GRA	m	l	13'537'909	0.63	283'001	26.7
Ga_lib_27	CATGA	Ga0122	NID	f	f	18'837'281	0.64	273'140	39.7
Ga_lib_27	GAGAT	Ga1163	BOH	m	l	17'815'656	0.53	286'542	29.3
Ga_lib_27	GTCAC	Ga0077	NID	f	f	13'431'633	0.59	273'935	25.9
Ga_lib_27	AGAGT	Ga0336	NID	m	p	33'777'653	0.60	277'001	64.9
Ga_lib_27	ATGCT	Ga0286	ROM	m	p	35'918'356	0.57	284'520	63.7
Ga_lib_27	TCGAG	Ga1079	GRA	m	f	11'471'522	0.61	285'134	21.8
Ga_lib_28	TAGCA	Ga1077	GRA	m	l	24'159'978	0.60	261'321	46.1
Ga_lib_28	GGAAAG	Ga1165	BOH	m	p	11'467'443	0.65	198'943	29.9
Ga_lib_28	TCGAG	Ga0082	NID	f	l	28'028'659	0.63	261'961	57.9
Ga_lib_28	GTCAC	Ga0099	ROM	m	f	20'048'656	0.66	245'333	45.6
Ga_lib_28	ACACG	Ga1431	GRA	f	l	7'167'002	0.65	151'191	24.0
Ga_lib_28	AGTCA	Ga0121	NID	m	l	24'008'515	0.60	260'951	46.2
Ga_lib_28	CATGA	Ga1118	UNT	f	f	16'869'124	0.55	256'994	29.2
Ga_lib_28	CGATA	Ga0293	ROM	f	f	36'304'322	0.59	267'543	68.2
Ga_lib_29	TCAGA	Ga1429	GRA	f	f	38'662	0.45	427	2.1

Ga_lib_29	TTCCG	Ga0157	ROM	m	f	30'258'306	0.49	285'983	45.3
Ga_lib_29	AGTCA	Ga1164	BOH	m	p	41'196'483	0.60	288'512	77.4
Ga_lib_29	ATGCT	Ga0160	NID	f	l	35'216'672	0.58	287'895	62.5
Ga_lib_29	CGATA	Ga0073	NID	f	p	34'578'247	0.57	286'229	62.1
Ga_lib_29	GAGAT	Ga1424	GRA	f	l	5'868'636	0.58	264'874	10.8
Ga_lib_29	GTCAC	Ga1420	GRA	f	l	6'131'229	0.55	264'470	10.6
Ga_lib_29	TGTGG	Ga1171	BOH	f	f	1'515'400	0.56	155'540	4.0
Ga_lib_30	TCGAG	Ga1125	UNT	m	f	61'344	0.70	6'371	2.5
Ga_lib_30	TTCCG	Ga1430	GRA	f	f	19'369'795	0.74	289'519	43.5
Ga_lib_30	ACACG	Ga1161	BOH	m	f	24'148'648	0.74	285'231	57.0
Ga_lib_30	AGTCA	Ga0313	NID	m	f	52'994'326	0.70	281'502	118.5
Ga_lib_30	GTCAC	Ga1425	GRA	f	l	28'300'487	0.71	283'839	63.3
Ga_lib_30	ATATC	Ga0233	NID	m	p	23'387'686	0.74	288'891	54.3
Ga_lib_30	CGATA	Ga1157	BOH	f	f	28'724'046	0.74	286'499	64.9
Ga_lib_31	TGACC	Ga1075	GRA	m	l	104'698	0.29	762	2.1
Ga_lib_31	TCGAG	Ga0081	NID	f	p	24'323'290	0.47	240'663	37.6
Ga_lib_31	ACTGC	Ga1162	BOH	m	f	4'933'837	0.56	226'296	9.3
Ga_lib_31	AGTCA	Ga0234	NID	m	l	22'305'914	0.54	252'085	40.0
Ga_lib_31	CATGA	Ga1418	GRA	f	l	4'870'606	0.58	229'118	9.4
Ga_lib_31	GAGAT	Ga0145	ROM	f	l	16'475'686	0.48	239'448	26.5
Ga_lib_31	GTCAC	Ga1099	UNT	m	f	14'728'702	0.56	257'067	26.8
Ga_lib_31	ATATC	Ga1169	BOH	m	f	3'332'506	0.59	206'238	6.9
Ga_lib_32	TGGTT	Ga1159	BOH	m	f	53'619	0.32	357	2.1
Ga_lib_32	TATAC	Ga1422	GRA	f	l	33'707'043	0.58	275'175	64.6
Ga_lib_32	CGATA	Ga0232	NID	m	p	23'694'642	0.46	273'891	35.6
Ga_lib_32	GAGAT	Ga1153	BOH	m	l	25'959'379	0.62	287'120	50.4
Ga_lib_32	GTCAC	Ga0084	NID	f	f	17'661'738	0.61	270'380	35.1
Ga_lib_32	ACTGC	Ga1073	GRA	m	p	7'825'103	0.64	273'226	15.7
Ga_lib_32	AGTCA	Ga1436	ROM	f	p	27'512'614	0.49	271'700	44.3
Ga_lib_32	CATGA	Ga1421	GRA	m	f	7'438'624	0.58	270'517	13.7

SUPPLEMENTARY REFERENCES

1. Moser, D., Roesti, M. & Berner, D. Repeated lake-stream divergence in stickleback life history within a Central European lake basin. *PLoS One* **7**, e50620 (2012)
2. Jones, F. C. et al. The genomic basis of adaptive evolution in threespine sticklebacks. *Nature* **484**, 55-61 (2012).

Chapter 7

Strong genome-wide divergence between sympatric
European river and brook lampreys

Strong genome-wide divergence between sympatric European river and brook lampreys

Catarina S. Mateus^{1,2,3,4,6,*}, Madlen Stange^{1,6}, Daniel Berner¹, Marius Roesti¹, Bernardo R. Quintella^{2,5}, M. Judite Alves⁴, Pedro R. Almeida^{2,3}, and Walter Salzburger^{1,*}

Lampreys, together with hagfishes, are the only extant representatives of jawless vertebrates and thus of prime interest for the study of vertebrate evolution [1]. Most lamprey genera occur in two forms with divergent life histories: a parasitic, anadromous and a non-parasitic, freshwater resident form [2–8]. The taxonomic status of such ‘paired species’ is disputed, however. While indistinguishable at larval stages, but clearly distinct as adults, they cannot be differentiated with available genetic data [6,7], which has fuelled speculations that the two forms may in fact represent products of phenotypic plasticity within a single species. Here, we use restriction site-associated DNA sequencing (RADseq) to examine the genetic population structure of sympatric European river (*Lampetra fluviatilis* L., 1758) and brook (*Lampetra planeri* Bloch, 1784) lampreys. We find strong genetic differentiation and identify numerous fixed and diagnostic single nucleotide polymorphisms (SNPs) between the two species, 12 of which can be unequivocally assigned to specific genes.

Lampreys — often referred to as cyclostomes because of their circular mouth — commonly occur as species pairs with distinct post-larval life histories. The so-called brook lampreys spend their entire life in freshwater, whereas their parasitic counterparts, the river lampreys, spend most of their adult life in the ocean or in estuaries and return to freshwater only for reproduction [2–8]. Whether these two forms are real species or are products of phenotypic plasticity in a single species has puzzled biologists for decades [2–4]. In the adult stage, river lampreys are much larger and morphologically distinct from brook

lampreys, which is why they have been described as distinct species. On the other hand, the larvae of the two forms are indistinguishable, the adults co-occur on breeding grounds and often spawn in common nests [8], and they produce viable offspring when crossed artificially [4], lending support to the plasticity hypothesis. Importantly, no genetic evidence is available to date that would suggest their separation (e.g., [6,7]). Sympatric European *L. fluviatilis* and *L. planeri* even share mitochondrial haplotypes, which was suggested to reflect ongoing gene flow or, alternatively, incomplete sorting of ancestral polymorphisms [6].

To address this ‘paired species’ conundrum in lampreys, we examined one pair in detail by means of Illumina-sequenced RAD. We considered 17 specimens of *L. fluviatilis* (Figure 1A) and 18 specimens of *L. planeri* (Figure 1B) collected from a common spawning site in the Sorraia River, a tributary of the Tagus River in Portugal, the southern limit of their distribution (see Supplemental Information published with the online version of this article). Sequences from one individual were used to build a pseudo-reference genome spanning 39,865 RAD loci (3.79 Mb), against which all individuals were aligned. Screening the alignments recovered 8,826 polymorphic RAD loci, yielding a total of 14,691 informative SNPs.

Global F_{ST} based on all SNPs between the two sympatric lampreys was no less than 0.37, suggesting strong genome-wide genetic differentiation despite the shared mitochondrial DNA haplotypes reported earlier for the exact same system [6]. Likewise, a genetic assignment test using Structure unambiguously separated the surveyed individuals into two distinct clusters (Figure 1C). The same result was obtained when the SNPs were analyzed in a phylogenetic context (Figure 1D). We thus provide the first genetic evidence for the taxonomic validity of the two European lamprey species *L. fluviatilis* and *L. planeri*. At the same time, we highlight the power of next generation sequencing technologies to resolve old questions in biology. Our data further agree with the assumption that resident lampreys are derived from migratory ones [2,3]. The genome scan revealed much greater genetic diversity in *L. fluviatilis* than in *L. planeri*. For instance, *L. fluviatilis*

displayed a 42% higher density of private SNPs than *L. planeri* (7,399 versus 5,198; binomial $p < 0.001$; see also branch-lengths in Figure 1D). In addition, the greater genetic diversity in the migratory species might also reflect the larger effective population size and less restricted gene flow. By contrast, we expect resident species to be more prone to genetic bottlenecks and genetic drift due to their reduced mobility.

To gain insight into genes potentially underlying the divergence between the sympatric lampreys, we screened the marker data for loci fixed for different alleles between the two species ($F_{ST} = 1$), identifying 166 such distinctive SNPs. Making use of the recently published genome of the sea lamprey [1], a distant relative of the species under investigation, we subjected these loci to reciprocal BLAST searches. This allowed us to link 12 of these loci to annotated genes. Interestingly, most of the genes showing fixed allelic differences between the two lampreys are related to functions that have previously been implicated in the adaptation to a migratory versus resident life-style in lampreys and bony fishes. For instance, fixed differences were found in the vasotocin gene, a major player in saltwater–freshwater osmoregulation and also involved in life history divergence [9], and in the gonadotropin-releasing hormone (GnRH), a key gene in gonadal development and differentiation [10]. We also found fixed genetic differences in four genes related to immune functions, three axial patterning genes, a pineal-gland-specific opsin, a sodium channel gene, and a tyrosine phosphatase gene. These genes are likely to contribute to ecologically based reproductive isolation in this lamprey system, paving the way for subsequent functional and evolutionary analyses. A more detailed discussion of the species-distinctive loci and their possible ecological role is provided in the Supplemental Information, along with a screen for large-scale genomic divergence between males and females in *L. planeri*.

In summary, we show that the sympatric lampreys *L. fluviatilis* and *L. planeri* are genetically highly distinct, and that the regions of strongest divergence contain several candidate genes for adaptation to a migratory versus resident life-style.



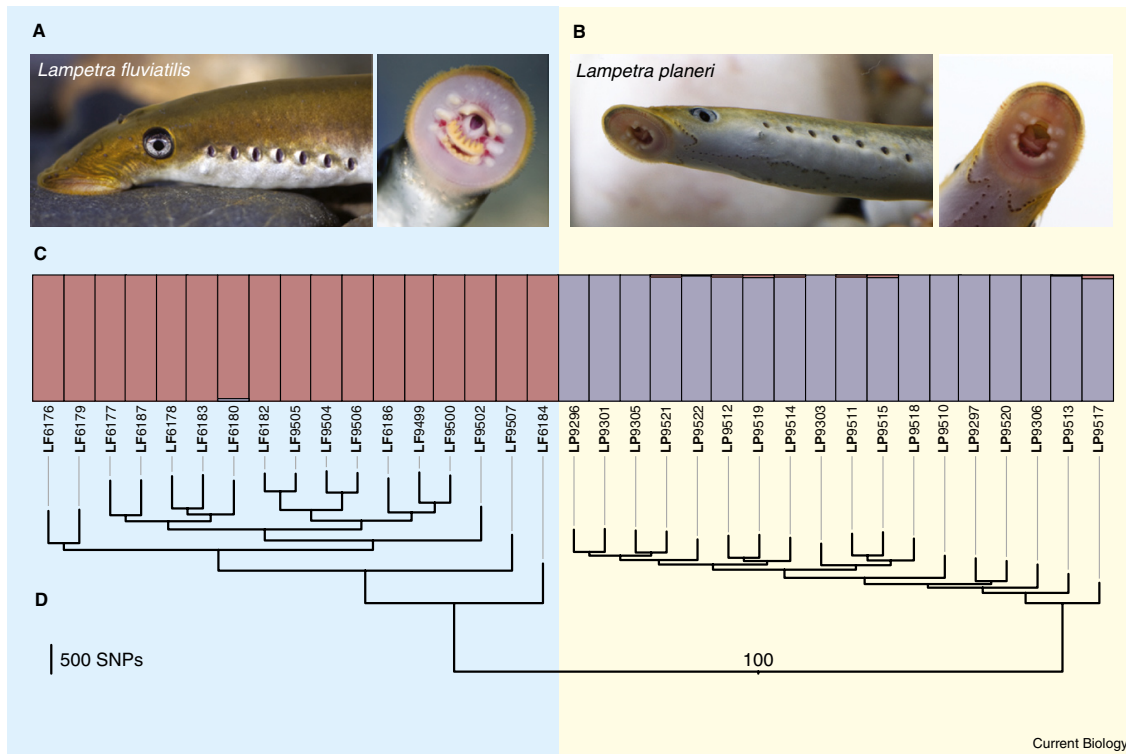


Figure 1. Genetic divergence in a lamprey species pair. The European river lamprey (*Lampetra fluviatilis*) (A) and the brook lamprey (*Lampetra planeri*) (B) are morphologically distinct in the adult stage. (C) A Bayesian population assignment test with Structure and a subsequent evaluation with Structure Harvester reveal the existence of two clusters ($K = 2$) in our SNP dataset, corresponding to the two sympatric species *L. fluviatilis* and *L. planeri*. Each bar represents the assignment probability (0 to 1) of a single specimen to one of these two clusters (color coded in red and purple, respectively). (D) Phylogeny of the 35 lamprey specimens from the Sorraia River in Portugal based on 14,691 SNPs and maximum parsimony in PAUP* (heuristic search with stepwise addition, TBR branch swapping and allowing polymorphisms). The specimens are grouped into two clades, which exactly match the two species *L. fluviatilis* and *L. planeri* (the bootstrap value for the basal branch is provided).

Supplemental Information

Supplemental Information including experimental procedures, a figure and a table can be found with this article online at <http://dx.doi.org/10.1016/j.cub.2013.06.026>.

Acknowledgments

This work was funded by the Fundação para a Ciência e a Tecnologia, Portugal (PhD grant SFRH/BD/44029/2008) and the European Science Foundation (ESF) exchange grant "Conservation Genomics: Amalgamation of Conservation Genetics and Ecological and Evolutionary Genomics" to CSM; the Swiss National Science Foundation (SNF; Ambizione grant PZ00P3_126391/1) to DB; and the University of Basel, the Swiss National Science Foundation (SNF; Sinergia grant CRSII3_136293), and the European Research Council (ERC) to WS. We are grateful to Brigitte Aeschbach and Nicolas Boileau for facilitating wet laboratory work, to I. Nissen and C. Beisel at the D-BSE Basel for Illumina sequencing, to Novocraft for sharing their aligner, and to

three anonymous reviewers for constructive comments.

References

- Smith, J.J., Kuraku, S., Holt, C., Sauka-Spengler, T., Jiang, N., Campbell, M.S., Yandell, M.D., Manousaki, T., Meyer, A., Bloom, O.E., et al. (2013). Sequencing of the sea lamprey (*Petromyzon marinus*) genome provides insight into vertebrate evolution. *Nat. Genet.* 45, 415–421.
- Hubbs, C.L. (1925). The life cycle and growth of lampreys. *Pap. Mich. Acad. Sci.* 4, 587–603.
- Hubbs, C.L. (1940). Speciation of fishes. *Am. Nat.* 74, 198–211.
- Enequist, P. (1937). Das Bachneunauge als Ökologische Modifikation des Flussneunauges. Über die Fluss- und Bachneunaugen Schwedens; vorläufige Mitteilung. *Ark. Zool.* 29, 1–22.
- Zanandrea, G. (1959). Speciation among lampreys. *Nature* 184, 380.
- Espanhol, R., Almeida, P.R., and Alves, M.J. (2007). Evolutionary history of lamprey paired species *Lampetra fluviatilis* (L.) and *Lampetra planeri* (Bloch) as inferred from mitochondrial DNA variation. *Mol. Ecol.* 16, 1909–1924.
- Docker, M.F., Mandrak, N.E., and Heath, D.D. (2012). Contemporary gene flow between "paired" silver (*Ichthyomyzon unicuspis*) and northern brook (*I. fossor*) lampreys: implications for conservation. *Cons. Gen.* 13, 823–835.
- Lasne, E., Sabatié, M.-R., and Evanno, G. (2010). Communal spawning of brook and river lampreys

(*Lampetra planeri* and *L. fluviatilis*) is common in the Oir River (France). *Ecol. Freshw. Fish* 19, 323–325.

- Balment, R.J., Lu, W., Weybourne, E., and Wame, J.M. (2006). Arginine vasotocin a key hormone in fish physiology and behaviour: a review with insights from mammalian models. *Gen. Comp. Endocrin.* 147, 9–16.
- Sower, S.A., and Kawachi, H. (2001). Update: brain and pituitary hormones of lampreys. *Comp. Biochem. Physiol. B* 129, 291–302.

¹Zoological Institute, University of Basel, Vesalgasse 1, CH-4051 Basel, Switzerland.

²Center of Oceanography, Faculty of Sciences of the University of Lisbon, Campo Grande, 1749-016 Lisbon, Portugal. ³Department of Biology, School of Sciences and Technology, University of Évora, Largo dos Colegiais 2, 7004-516, Évora, Portugal. ⁴National Museum of Natural History and Science & Center for Environmental Biology, University of Lisbon, Rua da Escola Politécnica 56/58, 1250-102 Lisbon, Portugal. ⁵Department of Animal Biology, Faculty of Sciences of the University of Lisbon, Campo Grande, 1749-016 Lisbon, Portugal.

⁶These authors contributed equally to this work.

*E-mail: csmateus@fc.ul.pt, walter.salzburger@unibas.ch

Chapter 7

Strong genome-wide divergence between sympatric
European river and brook lampreys

Supplementary material

Supplemental information

Strong genome-wide divergence between sympatric European river and brook lampreys

Catarina S. Mateus, Madlen Stange, Daniel Berner, Marius Roesti, Bernardo R. Quintella, M. Judite Alves, Pedro R. Almeida, and Walter Salzburger

Supplemental Inventory

Supplemental Figures and Tables

Figure S1

Table S1

Supplemental Experimental Procedures

Supplemental References

Supplemental Figures and Tables

Figure S1. Analysis of genomic divergence between males and females based on sex-specific read coverage across RAD loci in the lamprey *L. planeri* (A) and in threespine stickleback (*Gasterosteus aculeatus*) (B).

The existence of a relatively large genomic region highly differentiated between males and females will cause RAD loci within these regions to show sex-biased read coverage (details in [S1]). In a male-heterogametic system, for instance, read coverage for X-linked loci will be twofold higher in females than males as compared to autosomal loci for which read coverage between the sexes should be equal. The reason is that Y-linked sequences align poorly to their X-counterpart. Exactly this situation is found in stickleback: while most data points lie within the region predicted for autosomal loci (shown as yellow line in the plot), an additional cluster is visible along the line predicted for X-linked loci (green line; the expectation for W-linked loci in a female-heterogametic system is shown as blue line). By contrast, no deviation from the autosomal expectation is evident in *L. planeri*, indicating the absence of physically extensive genomic differentiation between males and females. Hence, if sex determination in this lamprey species is genetically based, the underlying system evolved without major chromosome divergence. Alternatively, sex determination might be under strong environmental influence, as generally assumed to occur in lampreys [S2–S4].

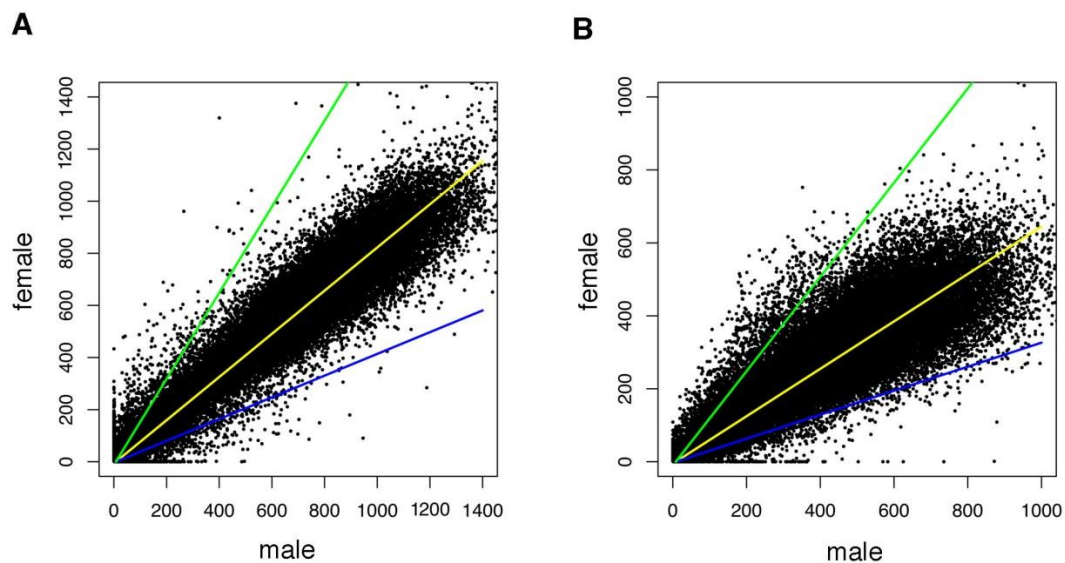


Table S1. Genes and gene families attained after BLAST of the SNPs with $F_{ST} = 1$.

Gene/Gene family	Function	References
Neurohypophysial gene (vasotocin)	Osmoregulation	[S5-S8]
Gonadotrophin-releasing hormone 2 precursor (GnRH2)	Gonadal maturation and migratory behavior	[S8-S11]
Pineal gland-specific opsin gene (P opsin)	Photoreception	[S12-S17]
Mannose-binding lectin-associated serine protease-1 (MASP-1) gene	Immunity	[S18, S19]
Ikaros-like genes (IKLF2)	Immunity	[S20, S21]
Variable lymphocyte receptor (VLR) gene	Immunity	[S22-S25]
CD45 gene (PTPRC, Protein tyrosine phosphatase, receptor type C)	Immunity	[S26, S27]
Homeobox genes (HoxW10a, Hox7, Emx)	Axial patterning and segmental identity	[S28-S31]
Voltage-gated sodium channel gene	Conduction of electrical signaling in nerves and muscles	[S32, S33]
Protein tyrosine phosphatase receptor type A precursor (PTPRA) gene	Regulation of cellular processes	[S5, S34]

Supplemental Experimental Procedures

Sampling

We collected (by electric fishing) 17 juvenile specimens of the anadromous *L. fluviatilis* at the start of their downstream trophic migration in January of two consecutive years (2009 and 2010), and 18 adult specimens of the resident *L. planeri* during the breeding season between late November 2009 and January 2010. All samples were collected in the Sorraia River, a tributary of the left bank of the Tagus River basin, where both species occur in sympatry. On the Iberian Peninsula, Tagus is the only river where the anadromous *L. fluviatilis* is known to occur, and it represents the southern range limit of both species [S35]. Tissue samples were preserved in 100% ethanol and deposited in the zoological collection ‘Museu Bocage’ of the Museu Nacional de História Natural e da Ciência (MUHNAC) (Lisbon, Portugal). Sampling was performed under the permission of the Instituto da Conservação da Natureza e das Florestas.

Restriction-site associated DNA (RAD) library preparation

RAD library preparation followed the protocol of Baird *et al.* [S36] and further modifications [S37, S38]. Briefly, DNA was extracted with the ‘DNeasy Blood & Tissue Kit’ (Qiagen) following the manufacturer’s protocol. Genomic DNA from each individual was digested with the *Sbf1* restriction enzyme. Each digest was then 5-mer barcoded for sample identification, and the 35 total samples were multiplexed into a single library. Final PCR enrichment was performed in 8 separate reactions to reduce amplification bias. Finally, the library was single-end sequenced with 100 cycles in a single Illumina HiSeq 2000 genome analyzer lane at D-BSSE Basel. Illumina reads are available from the Sequence Read Archive (SRA) at NCBI under the accession number PRJNA206554.

Marker generation

The reads were first quality-filtered and demultiplexed according to the individual barcodes. Using sequence data from the one individual with the highest read number, the reads were clustered by tolerating a maximum of two mismatches. For each cluster (representing a RAD locus), the consensus sequence was derived, and the unique consensus sequences were concatenated to form a 3.79 Mb pseudo-reference genome. These steps were carried out using Stacks v0.9996 [S39]. Next, data from each of the 35 individuals were aligned against the pseudo-reference genome using Novoalign v2.08.03 (<http://www.novocraft.com>), tolerating approximately six high-quality mismatches (-t ‘flag’ = 180). We enforced unique alignment, thereby avoiding that distinct loci in the pseudo-genome actually derived from the same locus in the true genome because of substantial polymorphisms. The alignments were then converted to *bam* format using Samtools v0.1.18 [S40]. Next, each RAD locus was genotyped at the whole-haplotype level. We here called a homozygous genotype when the dominant haplotype occurred in at least 18 copies and the second most frequent haplotype occurred less than six times. A heterozygote was called when the

two most frequent haplotypes occurred in at least 18 copies each. A locus not matching these criteria received a haploid genotype based on the dominant haplotype if that haplotype occurred in at least six copies, or were scored as missing data otherwise. As genotyping used fixed coverage thresholds, loci with excessive read coverage were down-sampled at random to 70x before genotyping (average read coverage per individual and RAD locus was 114.2, $sd = 59.8$). Finally, we combined the consensus sequences of all individuals to screen each RAD locus for SNPs. To exclude polymorphisms with low information content and technical artifacts [S41], SNPs displaying a minor allele frequency of 0.06 or lower were excluded from the data set. The resulting SNP panel for analysis included 34,267 SNPs. Genotyping and SNP calling was carried out using the R language [S42], benefiting from the bioconductor packages Biostrings and Rsamtools.

Population genetic and phylogenetic analyses

Prior to the analyses of genetic differentiation we eliminated SNPs with insufficient representation across individuals (threshold: 15 nucleotide calls from each population). The SNPs were used to calculate the haplotype-based fixation index (F_{ST}) [see S38] between the two samples. We then used Structure 2.3.4 [S43] to determine the number of genetic clusters (K) in our dataset and to estimate, for each individual, the assignment probability to these clusters. First, structure was run for 100,000 generations, with a burnin of 10,000 generation, and applying the admixture model for $K = 1$ to $K = 5$ and three independent replicates for each K . Using Structure Harvester [S44], we found that the most likely number of K was 2. We then repeated the Structure analysis for $K = 2$, running it for 500,000 generations (Figure 1C) and applying a burnin of 50,000. PAUP* [S45] was used to perform a phylogenetic analysis with the SNP dataset under maximum parsimony applying a heuristic search (stepwise addition and TBR branch swapping and allowing polymorphisms). Confidence assessment was performed with a bootstrap analysis and 1000 replicates. The resulting tree (Figure 1D) had a length of 22,632 steps. We also performed a neighbor-joining tree search (not shown), which produced a highly similar topology.

Screening fixed polymorphisms for candidate genes

For the 166 SNPs fixed for different alleles ($F_{ST} = 1$) between the samples, a homology search was first completed by performing a BLAST [S46] search on the NCBI public database. BLAST hits were then mapped to annotated genes in the Ensembl database [S47] making use of the recently released genome of the sea lamprey (*Petromyzon marinus*) [S48]. The hits were then confirmed by a reciprocal BLAST search, i.e., blasting the respective sea lamprey contig against all RAD tags. In total, we could link twelve RAD loci to annotated genes (Table S1). We found fixed differences in vasotocin, which is involved in many aspects of fish physiology and behavior, including circadian and seasonal biology, metabolism, reproduction and osmoregulation [S5-S8]; in the gonadotropin-releasing hormone 2 (GnRH2), a key gene in gonadal development and differentiation, and regulation of the reproductive and migratory behavior, by controlling secretion of pituitary hormones [S8-S11]; in

the non-visual pineal gland-specific opsin gene (P opsin), which is key in photoreception in lamprey larvae, controlling the changes in body coloration and metamorphosis, and in adults through control of sexual maturation [S12-S17]. We found four genes implicated with immune functions: a mannose-binding lectin-associated serine protease (MASP), the ikaros factor-like 2 gene (IKFL2), variable lymphocyte receptor (VLR), and the protein tyrosine phosphatase receptor type C (PTPRC or CD45) [see S18-S27]. We also found hits with three homeobox genes (HoxW10a, Hox7, Emx), which are known to be involved in the specification and patterning of different regions along the body axes [S28-S31]. In particular, Emx is known to play a major role in forebrain development. Hits were also found with the voltage-gated sodium channel gene, known to play an essential role in physiology through the initiation and propagation of action potentials in neurons and other electrically excitable cells such as myocytes and endocrine cells [S32, S33], and finally, in the protein tyrosine phosphatase receptor type A precursor (PTPRA). The protein encoded by PTPRA is a member of the protein tyrosine phosphatase (PTPase) family. PTPases are involved in a variety of cellular processes including cell activation, growth and differentiation, mitotic cycle, and oncogenic transformation [S5, S34].

Genomic screen for large sex-specific regions

We here used a subsample of five females and seven males from the resident species *L. planeri*. This included all lamprey individuals for which sex was known (note that *L. fluviatilis* were sampled as migrating juveniles, precluding the phenotypic identification of sex). The full alignments of these 12 individuals were used to screen visually for the presence of a major sex-linked genomic region (Figure S1). For this, the total number of reads was counted separately across all males and all females at each of the 38,308 total RAD loci. For each locus, the total female count was then plotted against the total male count. The rationale was that RAD loci in sex-specific regions should exhibit systematic read coverage bias between males and females relative to loci in autosomal regions, because of differential alignment success to the reference sequence [for details see S1]. This approach should thus allow detecting at least large-scale differentiation between males and females visually. For comparison, we performed an analogous investigation with exactly the same sample size using RAD data from threespine stickleback [S1], a species with a major XY chromosomal system [S48].

Supplemental References

- S1. Roesti, M., Moser, D., and Berner, D. (2013). Recombination in the threespine stickleback genome - patterns and consequences. *Mol. Ecol.* *22*, 3014-3027.
- S2. Beamish, F. W. H. (1993). Environmental sex determination in southern brook lamprey, *Ichthyomyzon gagei*. *Can. J. Fish. Aquat. Sci.* *50*, 1299–1307.
- S3. Lowartz, S. M., and Beamish, F. W. H. (2000). Novel perspectives in sexual lability through gonadal biopsy in larval sea lampreys. *J. Fish Biol.* *56*, 743–757.
- S4. Docker, M. F., and Beamish, F. W. H. (1994). Age, growth, and sex ratio among populations of least brook lamprey, *Lampetra aepyptera*, larvae: an argument for environmental sex determination. *Env. Biol. Fishes* *41*, 191–205.
- S5. Gwee, P.-C., Tay, B.-H., Brenner, S., and Venkatesh, B. (2009). Characterization of the neurohypophysial hormone gene loci in elephant shark and the Japanese lamprey: origin of the vertebrate neurohypophysial hormone genes. *BMC Evol. Biol.* *9*, 47.
- S6. Balment, R. J., Warne, J. M., Tierney, M., and Hazon, N. (1993). Arginine vasotocin and fish osmoregulation. *Fish Physiol. Biochem.* *11*, 189–194.
- S7. Balment, R. J., Lu, W., Weybourne, E., and Warne, J. M. (2006). Arginine vasotocin a key hormone in fish physiology and behaviour: a review with insights from mammalian models. *Gen. Compar. Endocrinol.* *147*, 9–16.
- S8. Sower, S. A., and Kawachi, H. (2001). Update: brain and pituitary hormones of lampreys. *Comp. Biochem. Physiol. B* *129*, 291–302.
- S9. Onuma, T., Higa, M., Ando, H., Ban, M., and Urano, A. (2005). Elevation of gene expression for salmon gonadotropin-releasing hormone in discrete brain loci of prespawning chum salmon during upstream migration. *J. of Neurobiol.* *63*, 126–145.
- S10. Gazourian, L., Deragon, K. L., Chase, C. F., Pati, D., Habibi, H. R., and Sower, S. A. (1997). Characteristics of GnRH binding in the gonads and effects of lamprey GnRH-I and -III on reproduction in the adult sea lamprey. *Gen. Compar. Endocrinol.* *108*, 327–339.
- S11. Rissman, E. F. (1996). Behavioral regulation of gonadotropin-releasing hormone. *Biol. of Reproduct.* *54*, 413–419.
- S12. Joss, J. M. P. (1973). Pineal-gonad relationships in the lamprey *Lampetra fluviatilis*. *Gen. Compar. Endocrinol.* *21*, 118–122.
- S13. Joss, J. M. P. (1973). The pineal complex, melatonin, and color change in the lamprey *Lampetra*. *Gen. Compar. Endocrinol.* *21*, 188–195.
- S14. Cole, W. C., and Youson, J. H. (1981). The effect of pinealectomy, continuous light, and continuous darkness on metamorphosis of anadromous sea lampreys, *Petromyzon marinus* L. *J. Exp. Zool.* *218*, 397–404.
- S15. Pu, G. A., and Dowling, J. E. (1981). Anatomical and physiological characteristics of pineal photoreceptor cell in the larval lamprey, *Petromyzon marinus*. *J. Neurophysiol.* *46*, 1018–1038.
- S16. Tamotsu, S., and Morita, Y. (1986). Photoreception in pineal organs of larval and adult lampreys, *Lampetra japonica*. *J. Comp. Physiol. A* *159*, 1–5.
- S17. Yokoyama, S., and Zhang, H. (1997). Cloning and characterization of the pineal gland-specific opsin gene of marine lamprey (*Petromyzon marinus*). *Gene* *202*, 89–93.
- S18. Endo, Y., Takahashi, M., Nakao, M., Saiga, H., Sekine, H., Matsushita, M., Nonaka, M., and Fujita, T. (1998). Two lineages of mannose-binding lectin-

- associated serine protease (MASP) in vertebrates. *J. Immunol.* *161*, 4924–4930.
- S19. Endo, Y., Nonaka, M., Saiga, H., Kakinuma, Y., Matsushita, A., Takahashi, M., Matsushita, M., and Fujita, T. (2003). Origin of mannose-binding lectin-associated serine protease (MASP)-1 and MASP-3 involved in the lectin complement pathway traced back to the invertebrate, amphioxus. *J. Immunol.* *170*, 4701–4707.
- S20. Mayer, W. E., Huigin, C. O., Tichy, H., Terzic, J., and Saraga-Babic, M. (2002). Identification of two Ikaros-like transcription factors in lamprey. *Scan. J. Immunol.* *55*, 162–170.
- S21. Haire, R. N., Miracle, A. L., Rast, J. P., and Litman, G. W. (2000). Members of the Ikaros gene family are present in early representative vertebrates. *J. Immunol.* *165*, 306–312.
- S22. Cooper, M. D., and Alder, M. N. (2006). The evolution of adaptive immune systems. *Cell* *124*, 815–822.
- S23. Pancer, Z., Saha, N. R., Kasamatsu, J., Suzuki, T., Amemiya, C. T., Kasahara, M., and Cooper, M. D. (2005). Variable lymphocyte receptors in hagfish. *Proc. Natl. Acad. Sci. USA* *102*, 9224–9229.
- S24. Alder, M. N., Rogozin, I. B., Iyer, L. M., Glazko, G. V, Cooper, M. D., and Pancer, Z. (2005). Diversity and function of adaptive immune receptors in a jawless vertebrate. *Science* *310*, 1970–1973.
- S25. Boehm, T., McCurley, N., Sutoh, Y., Schorpp, M., Kasahara, M., and Cooper, M. D. (2012). VLR-based adaptive immunity. *Ann. Rev. Immunol.* *30*, 203–220.
- S26. Uinuk-ool, T., Nikolaidis, N., Sato, A., Mayer, W. E., and Klein, J. (2005). Organization, alternative splicing, polymorphism, and phylogenetic position of lamprey CD45 gene. *Immunogenetics* *57*, 607–617.
- S27. Uinuk-ool, T., Mayer, W. E., Sato, A., Dongak, R., Cooper, M. D., and Klein, J. (2002). Lamprey lymphocyte-like cells express homologs of genes involved in immunologically relevant activities of mammalian lymphocytes. *Proc. Natl. Acad. Sci. USA* *99*, 14356–14361.
- S28. Irvine, S. Q., Carr, J. L., Bailey, W. J., Kawasaki, K., Shimizu, N., and Amemiya, C. T. (2002). Genomic analysis of Hox clusters in the sea lamprey *Petromyzon marinus*. *J. Exp. Zool. B* *294*, 47–62.
- S29. Pendleton, J. W., Nagai, B. K., Murtha, M. T., and Ruddle, F. H. (1993). Expansion of the Hox gene family and the evolution of chordates. *Proc. Natl. Acad. Sci. USA* *90*, 6300–6304.
- S30. Boncinelli, E. (1999). Otx and Emx Homeobox Genes in Brain Development. *The Neuroscientist* *5*, 164–172.
- S31. Tank, E. M., Dekker, R. G., Beauchamp, K., Wilson, K. A., Boehmke, A. E., and Langeland, J. A. (2009). Patterns and consequences of vertebrate Emx gene duplications. *Evol. Devel.* *11*, 343–353.
- S32. Yu, F. H., and Catterall, W. A. (2003). Overview of the voltage-gated sodium channel family. *Genome Biol.* *4*, 207.
- S33. Novak, A. E., Jost, M. C., Lu, Y., Taylor, A. D., Zakon, H. H., and Ribera, A. B. (2006). Gene duplications and evolution of vertebrate voltage-gated sodium channels. *J. Mol. Evol.* *63*, 208–21.
- S34. Saadat, M., Nakamura, K., Mizuno, Y., Kikuchi, K., and Yoshida, M. C. (1995). Regional localization of rat and mouse protein-tyrosine phosphatase

- PTP α /LRP gene (Ptp α) by fluorescence *in situ* hybridization. *Japan. J. Genetics* 70, 669–674.
- S35. Mateus, C. S., Rodríguez-Muñoz, R., Quintella, B. R., Alves, M. J., and Almeida, P. R. (2012). Lampreys of the Iberian Peninsula: distribution, population status and conservation. *Endangered Species Research* 16, 183–198.
- S36. Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Zachary, A., Selker, E. U., Cresko, W. A., and Johnson, E. A. (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE* 3, e3376.
- S37. 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 Genetics* 6, e1000862.
- S38. Roesti, M., Hendry, A. P., Salzburger, W., and Berner, D. (2012). Genome divergence during evolutionary diversification as revealed in replicate lake-stream stickleback population pairs. *Mol. Ecol.* 21, 2852–62.
- S39. Catchen, J. M., Amores, A., Hohenlohe, P., Cresko, W., and Postlethwait, J. H. (2011). Stacks: building and genotyping Loci de novo from short-read sequences. *G3* 1, 171–182.
- S40. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)* 25, 2078–2079.
- S41. Roesti, M., Salzburger, W., and Berner, D. (2012). Uninformative polymorphisms bias genome scans for signatures of selection. *BMC Evol. Biol.* 12, 94.
- S42. R Development Core Team (2010). R: A language and environment for statistical computing (Vienna, Austria: R Foundation for Statistical Computing).
- S43. Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multicolour genotype data. *Genetics* 155, 945–959.
- S44. Earl, D. A. and vonHoldt, B. M. (2012). Structure harvester: a website and program for visualizing Structure output and implementing the Evanno method. *Cons. Genetic Res.* 4, 359–361.
- S45. Swofford, D. L. (2003). PAUP* - Phylogenetic Analyses Using Parsimony and other methods, version 4.0. Sinauer, Sunderland, MA.
- S46. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic Local Alignment Search Tool. *J. Mol. Biol.* 215, 403–410.
- S47. Flicek, P., Ahmed, I., Amode, M. R., Barrell, D., Beal, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fairley, S., et al. (2013). Ensembl 2013. *Nucl. Acids Res.* 41, D48–D55.
- S48. Smith, J. J., Kuraku, S., Holt, C., Sauka-Spengler, T., Jiang, N., Campbell, M. S., Yandell, M. D., Manousaki, T., Meyer, A., Bloom, O. E., et al. (2013). Sequencing of the sea lamprey (*Petromyzon marinus*) genome provides
- S49. Peichel, C. L., Ross, J. A., Matson, C. K., Dickson, M., Grimwood, J., Schmutz, J., Myers, R. M., Mori, S., Schluter, D., and Kingsley, D. M. (2004). The master sex-determination locus in threespine sticklebacks is on a nascent Y chromosome. *Curr. Biol.* 14, 1416–1424.

Chapter 8

Natural selection: it's a many-small world after all

the sparse code extends into the time domain, in this case with a relevant time resolution of tens to hundreds of milliseconds. Extending into the time domain naturally increases the coding capacity at this stage of processing.

The Gupta and Stopfer paper [13] is also important because it sheds light on the read-out of sparse representations. While we have a reasonable understanding of sparse representations in the insect mushroom bodies and the vertebrate cortex, the read-out is still poorly understood. The authors clearly show that the representation changes from dense in the sensory system to sparse in the Kenyon cells and again to dense at the level of the output neurons. Changing coding schemes might be a common principle, because recent work in the mammalian cortex has shown that sparse representation in cortical input layers is transformed to a dense representation in output layers (for review see [4,17]). Notably, both cortex and insect mushroom bodies are involved in associative learning and theoretical studies have shown that sparse representations improve learning of associative representations (for example, [18,19]).

While the precise role of the mushroom body output neurons is currently not clear, it is unlikely that they constitute a 'simple' continuation of the olfactory pathway providing just another olfactory code. The mushroom bodies are centers for multimodal processing and associative memory, and reward-based mechanisms of plasticity have been shown in the synapses between Kenyon cells and output neurons [20]. Thus, the output neurons might be involved in recoding sensory representations to an experience-dependent value code that represents the behavioral relevance of sensory input. This notion would be in line with previous work, which found little odor identity coding, but strong odor-reward association encoding after memory consolidation at the mushroom body output [14]. A rapid representation of the behaviorally relevant stimuli might be a prerequisite for behavioral decision making based on experience-dependent memory.

While this new study [13] shows the importance of the time domain

for sparse coding in biological systems, this concept might also be inspiring for computer science. In the field of machine learning high-dimensional sparse projections of inputs are used to improve stimulus classification with reinforcement learning. Since this analogy between sparse coding in biological systems and in machine learning has been repeatedly outlined (for example, [18]), it might be of interest to better explore temporal coding schemes for machine learning algorithms, for example, in order to increase the capacity of artificial object recognition systems.

References

1. Barlow, H.B. (1969). Trigger features, adaptation and economy of impulses. In *Information Processing in the Nervous System* (pp. 209–230). Springer Berlin Heidelberg.
2. Vinje, W.E., and Gallant, J.L. (2000). Sparse coding and decorrelation in primary visual cortex during natural vision. *Science* 287, 1273–1276.
3. Hromádka, T., DeWeese, M.R., and Zador, A.M. (2008). Sparse representation of sounds in the unanesthetized auditory cortex. *PLoS Biol.* 6, e16.
4. Wolfe, J., Houweling, A.R., and Brecht, M. (2010). Sparse and powerful cortical spikes. *Curr. Opin. Neurobiol.* 20, 306–312.
5. Isaacson, J.S. (2010). Odor representations in mammalian cortical circuits. *Curr. Opin. Neurobiol.* 20, 328–331.
6. Heisenberg, M. (2003). Mushroom body memoir: from maps to models. *Nat. Rev. Neurosci.* 4, 266–275.
7. Menzel, R. (2012). The honeybee as a model for understanding the basis of cognition. *Nat. Rev. Neurosci.* 13, 758–768.
8. Perez-Orive, J., Mazor, O., Turner, G.C., Cassenaer, S., Wilson, R.I., and Laurent, G. (2002). Oscillations and sparsening of odor representations in the mushroom body. *Science* 297, 359–365.
9. Ito, I., Ong, R.C.Y., Raman, B., and Stopfer, M. (2008). Sparse odor representation and olfactory learning. *Nat. Neurosci.* 11, 1177–1184.
10. Demmer, H., and Kloppenburg, P. (2009). Intrinsic membrane properties and inhibitory synaptic input of Kenyon cells as mechanisms for sparse coding? *J. Neurophys.* 102, 1538–1550.
11. Honegger, K.S., Campbell, R.A., and Turner, G.C. (2011). Cellular-resolution population imaging reveals robust sparse coding in the *Drosophila* mushroom body. *J. Neurosci.* 31, 11772–11785.
12. Farkhooi, F., Froese, A., Müller, E., Menzel, R., and Nawrot, M.P. (2013). Cellular adaptation facilitates sparse and reliable coding in sensory pathways. *PLoS Comp. Biol.* 9, e1003251.
13. Gupta, N., and Stopfer, M. (2014). A temporal channel for information in sparse sensory coding. *Curr. Biol.* 24, 2247–2256.
14. Strube-Bloss, M.F., Nawrot, M.P., and Menzel, R. (2011). Mushroom body output neurons encode odor-reward associations. *J. Neurosci.* 31, 3129–3140.
15. Gupta, N., and Stopfer, M. (2012). Functional analysis of a higher olfactory center, the lateral horn. *J. Neurosci.* 32, 8138–8148.
16. Güting, R. (2014). To spike, or when to spike? *Curr. Opin. Neurobiol.* 25, 134–139.
17. Harris, K.D., and Mrcic-Flögel, T.D. (2013). Cortical connectivity and sensory coding. *Nature* 503, 51–58.
18. Huerta, R., and Nowotny, T. (2009). Fast and robust learning by reinforcement signals: explorations in the insect brain. *Neural Comput.* 21, 2123–2151.
19. Palm, G. (2013). Neural associative memories and sparse coding. *Neural Networks* 37, 165–171.
20. Cassenaer, S., and Laurent, G. (2012). Conditional modulation of spike-timing-dependent plasticity for olfactory learning. *Nature* 482, 47–52.

¹Biocenter, Institute for Zoology, Cologne Excellence Cluster on Cellular Stress Responses in Aging Associated Diseases (CECAD), University of Cologne, 50674 Cologne, Germany. ²Theoretical Neuroscience, Institute of Biology, Freie Universität Berlin, 14195 Berlin, Germany. E-mail: peter.kloppenburger@uni-koeln.de, martin.nawrot@fu-berlin.de

<http://dx.doi.org/10.1016/j.cub.2014.08.041>

Natural Selection: It's a Many-Small World After All

Understanding adaptive phenotypic change and its genetic underpinnings is a major challenge in biology. Threespine stickleback fish, experimentally exposed to divergent semi-natural environments, reveal that adaptive diversification can happen readily, affects many traits and involves numerous genetic loci across the genome.

Marius Roesti* and Walter Salzburger

Populations exposed to contrasting environments typically become different in phenotype and may ultimately split into distinct, reproductively isolated species [1].

The genetic basis of phenotypic change during this process remains poorly understood. Major drawbacks are that most research focuses on a few traits in lab-reared specimens, targets phenotypes with a simple genetic architecture or uses indirect



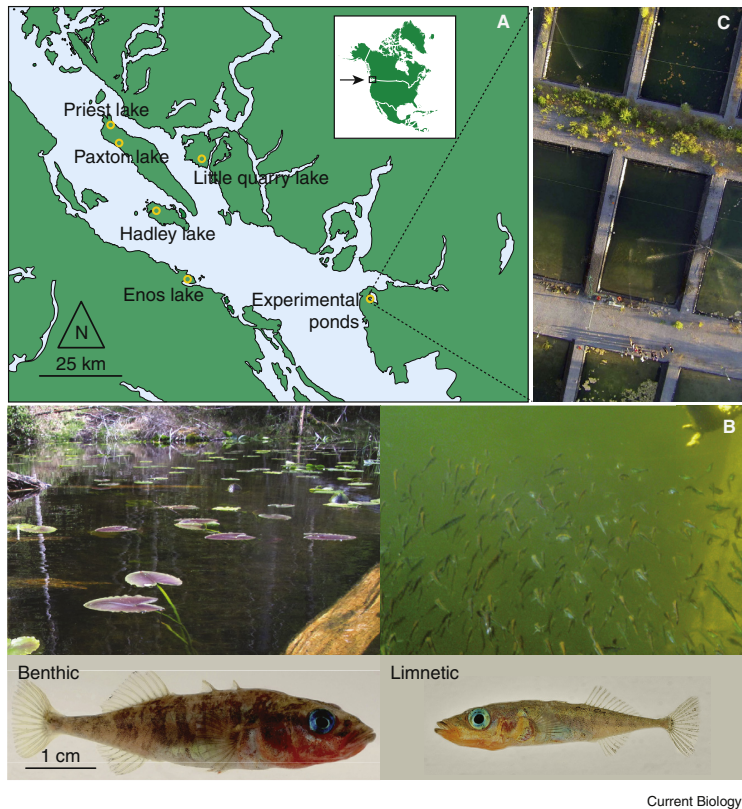


Figure 1. Co-occurring benthic and limnetic stickleback, and their natural and reconstructed habitats.

(A) Four benthic–limnetic species pairs have been officially listed by the Committee on the Status of Endangered Wildlife in Canada (COSEWIC) so far, of which Hadley Lake stickleback went extinct in the 1990s and the species pair in Enos Lake has collapsed into a hybrid swarm. Both these events are associated with invasive species. A fifth species pair was discovered in Little Quarry Lake in 2007. (B) The shallow-water benthic habitat of Paxton Lake, the contrasting deep open-water limnetic habitat (pictures courtesy of Jenny Boughman [benthic], Gina Conte [limnetic]), and a representative adult male stickleback from each habitat type (pictures courtesy of Gerrit Velema; note that the benthic specimen originates from close-by Priest Lake). (C) Bird’s eye view of the experimental pond facility at the University of British Columbia (Vancouver, Canada). Each pond measures 25 by 15 meters, holds 750,000 liters of water, and mimics a natural lake with both a benthic and limnetic zone (picture courtesy of Thor Veen).

inference from population genomic surveys that lack functional trait information. To better understand the genetics underlying ecological divergence, we should therefore investigate the complete set of traits that bring whole individuals closer to their performance optimum in a particular natural environment [2,3]. This task has proven extremely difficult, and uncovering the genetic basis of adaptation remains a challenge. In a recent study, Arnegard and co-workers [4] take on this challenge by using genetic mapping to

study niche divergence in threespine stickleback under semi-natural conditions.

Stickleback fish are an important model system for speciation research, especially in a few postglacial lakes in British Columbia, Canada, where both inshore (benthic) and offshore (limnetic) stickleback species (sometimes referred to as ‘ecomorphs’) have evolved repeatedly in less than 12,000 years (Figure 1A) [5]. These co-occurring species are reproductively isolated through different morphological, behavioral

and physiological adaptations to their contrasting habitats (Figure 1B). Hybrids with intermediate phenotypes are occasionally found, but their performance is relatively low in both habitats as compared to the pure species [6,7]. These aspects make benthic and limnetic stickleback one of the most explicit examples of natural selection’s predominant role in the origin of new species.

In their experiment, Arnegard *et al.* [4] released F1 hybrids from artificial crosses between benthic and limnetic stickleback collected from Paxton Lake (Figure 1A,B) into a large experimental pond (Figure 1C). This pond, which includes both shallow-water benthic and deep-water limnetic zones, approximates the distinct habitats the two species occupy in the wild. The authors allowed the F1s to mate freely and, after six months, sampled more than 600 second-generation F2 hybrids throughout the pond. Among these F2s, stable isotope profiles and stomach content analyses indicated extensive variation in niche exploitation along the benthic–limnetic axis. Importantly, hybrids at both ends of the benthic–limnetic diet spectrum (those close to pure species phenotypes) grew larger, suggesting that they performed better than other hybrids (those deviating from pure species phenotypes). Variation in niche use was further associated with functional and morphological divergence in their feeding apparatus and body shape.

To decipher the genetic architecture of this divergence, Arnegard *et al.* [4] used quantitative trait locus (QTL) mapping, an approach that provides a strong test for causality by linking phenotypic to genetic variation within an experimental cross population. The authors found that many loci across the stickleback’s genome, each with a small to moderate effect at the phenotype level, underlie benthic–limnetic divergence. Moreover, several QTLs contributed additively and more or less evenly to whole-organism niche performance. That is, the addition of a favorable allele at any of these QTLs brought an individual’s overall phenotype a similarly small step closer to its fitness optimum. In contrast to the well-adapted benthic and limnetic hybrids, F2 individuals with an intermediate diet signature were smaller, had a mixed combination of

benthic and limnetic alleles and were intermediate in phenotype. Finally, F2 individuals showing the strongest growth deficits exhibited conflicting combinations of the ecologically relevant traits, making them particularly maladapted for either of the two trophic habitats. Arnegard *et al.* [4] thus provide an elegant and rare demonstration for how variation in the genotype translates, through the phenotype, to fitness differences among individuals.

The study also confirms a general finding emerging from high-resolution genome scans between ecologically divergent populations [8–12]: adaptation is a complex process involving many genetic loci. A first reason is that adaptation is likely to require shifts at many phenotypic traits, including behavior, morphology, physiology and life history. A second reason is that even single ecologically relevant traits are commonly controlled by many genetic loci, each with a small phenotypic effect [2]. Although some traits certainly do have an underlying simple (nearly Mendelian) genetic basis [13,14], high-resolution sequencing technology has revealed that some of these ‘single locus with large phenotypic effect’ examples are in reality much more genetically complex than initially thought [15,16]. These insights raise an important question: to what extent are the few straightforward cases of genotype-to-phenotype relationships for single traits representative of adaptation’s complexity as a whole?

When studying something as complex as adaptation, it is essential to choose an appropriate methodology and to recognize its possible limitations [17]. For example, the crux with traditional QTL mapping is the focus on a few traits and only a single cross — that is, all F2 individuals derive from the same two grandparents. The genetic variation in such a cross does not capture the allelic richness available to selection in a natural population and is likely to limit the available phenotypic variation. Furthermore, most QTL studies cannot easily connect their results to the natural context (but see, e.g., [18,19]). Arnegard *et al.* [4] reduced these limitations by using semi-natural ponds and first-generation F1 hybrids from four independent crosses. In this way, instead of having a maximum of four allelic variants per locus, as is the case

in a single F2 QTL cross (two alleles from each grandparent), up to 16 possible variants were exposed to selection in their study. In addition, all individuals were free to choose their mating partners, habitat, and diet. This puts the study by Arnegard and colleagues [4] far beyond traditional QTL mapping. Most notably, the authors are able to link their phenotypic and genetic findings to adaptive population divergence, and hence, fitness consequences within distinct semi-natural habitats.

Nevertheless, some limitations associated with QTL mapping remain. The relatively low marker density used to genotype the individuals (less than 500 markers) and the constraints given by only a single generation of genetic admixture (from first to second generation hybrids) inevitably result in a relatively limited resolution when inferring genomic regions associated with phenotypic traits [2,17]. These limitations make it impossible to determine whether mapped genomic regions contain multiple close-by loci, each with a very small and possibly non-additive contribution to trait variation, or a single locus with a relatively larger phenotypic effect. Improvements could include sampling a QTL cross population after more generations, increasing marker resolution and adding association mapping in natural, highly variable populations. Even so, these approaches remain constrained to finding loci with relatively large phenotypic effects [20]. Also, because F2 hybrids were exposed to ecologically different habitats throughout their lives, some portion of their trait variation might reflect phenotypic plasticity, which could confound QTL inference. A solution here would be to re-map the focal traits in an F2-cross raised under the same standardized conditions. Finally, we need to establish to what extent our current methodological toolkit is biased towards detecting additive over more complex non-additive genetics [2].

Interestingly, the experiment also yielded an unpredicted outcome: the smallest F2 individuals, which showed mismatches in functional traits, were feeding on springtails, a food resource fortuitously abundant within the experimental pond but largely absent in the natural habitat. We can only speculate as to how this new resource

could have influenced evolution in the pond if the experiment had been run for more generations. Despite the availability of this alternative food type, the springtail-feeders might not persist through future generations. It is also possible, although rather unlikely, that this group becomes well-adapted to the new springtail-foraging niche, resulting in a brand new ecomorph next to the limnetic and benthic stickleback. Finally, these small intermediate phenotypes could facilitate gene flow between the benthics and limnetics, allowing some combinations of benthic and limnetic alleles to be relatively fit. This in turn might hinder further adaptive divergence between the pure ecomorphs and counteract any possible experimental speciation. The occurrence of the springtail-feeders shows how difficult it is to precisely reconstruct the ecological conditions shaping divergence in the wild. Furthermore, it highlights that learning about the predictability of evolution requires comparable and replicated studies, within and across organisms.

Overall, the study by Arnegard *et al.* [4] demonstrates that, despite involving many traits and loci, important fitness variation can emerge immediately when the right allelic variants are available to selection. Another interesting finding is that the genetic architecture underlying reduced environment-dependent hybrid viability and thus reproductive isolation might be largely additive. This contrasts with the idea of environment-independent reproductive isolation (i.e., due to intrinsic genetic incompatibilities) that is mainly caused by deleterious non-additive gene interactions. These exciting novel insights point to the future promise of taking experimental (genetic) approaches out into nature.

References

1. Schluter, D. (2000). *The Ecology of Adaptive Radiation* (Oxford: Oxford University).
2. Mackay, T.F.C., Stone, E.A., and Ayroles, J.F. (2009). The genetics of quantitative traits: challenges and prospects. *Nat. Rev. Genet.* 10, 565–577.
3. Barrett, R.D., and Hoekstra, H.E. (2011). Molecular spandrels: tests of adaptation at the genetic level. *Nat. Rev. Genet.* 12, 767–780.
4. Arnegard, M.E., McGee, M.D., Matthews, B., Marchinko, K.B., Conte, G.L., Kabir, S., Bedford, N., Bergek, S., Chan, Y.F., Jones, F.C., *et al.* (2014). Genetics of ecological divergence during speciation. *Nature* 511, 307–311.
5. Bell, M.A., and Foster, S.A. (1994). *The Evolutionary Biology of the Threespine Stickleback* (Oxford: Oxford University).
6. McPhail, J.D. (1992). *Ecology and Evolution of sympatric sticklebacks (Gasterosteus)*:

- evidence for a species pair in Paxton Lake, Texada Island, British Columbia. *Can. J. Zool.* 70, 361–369.
- Gow, J.L., Peichel, C.L., and Taylor, E.B. (2007). Ecological selection against hybrids in natural populations of sympatric threespine sticklebacks. *J. Evol. Biol.* 20, 2173–2180.
 - Hohenlohe, P.A., Bassemir, 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.
 - Lawniczak, M.K.N., Emrich, S.J., Holloway, A.K., Regier, A.P., Olson, M., White, B., Redmond, S., Fulton, L., Appelbaum, E., Godfrey, J., *et al.* (2010). Widespread divergence between incipient *Anopheles gambiae* species revealed by whole genome sequences. *Science* 330, 512–514.
 - Fournier-Level, A., Korte, A., Cooper, M.D., Nordborg, M., Schmitt, J., and Wilczek, A.M. (2011). A map of local adaptation in *Arabidopsis thaliana*. *Science* 334, 86–89.
 - Roesti, M., Hendry, A.P., Salzburger, W., and Berner, D. (2012). Genome divergence during evolutionary diversification as revealed in replicate lake-stream stickleback population pairs. *Mol. Ecol.* 21, 2852–2862.
 - Soria-Carrasco, V., Gompert, Z., Comeault, A.A., Farkas, T.E., Parchman, T.L., Johnston, J.S., Buerkle, C.A., Feder, J.L., Bast, J., *et al.* (2014). Stick insect genomes reveal natural selection's role in parallel speciation. *Science* 344, 738–742.
 - Shapiro, M.D., Marks, M.E., Peichel, C.L., Blackman, B.K., Nereng, K.S., Jonsson, B., Schluter, D., and Kingsley, D.M. (2004). Genetic and developmental basis of evolutionary pelvic reduction in threespine sticklebacks. *Nature* 428, 717–723.
 - Colosimo, P.F., Hosemann, K.E., Balabhadra, S., Villareal, G., Jr., Dickson, M., Grimwood, J., Schmutz, J., Myers, R.M., Schluter, D., and Kingsley, D.M. (2005). Widespread parallel evolution in sticklebacks by repeated fixation of ectodysplasin alleles. *Science* 307, 1928–1933.
 - Joron, M., Frezal, L., Jones, R.T., Chamberlain, N.L., Lee, S.F., Haag, C.R., Whibley, A., Becuwe, M., Baxter, S.W., Ferguson, L., *et al.* (2011). Chromosomal rearrangements maintain a polymorphic supergene controlling butterfly mimicry. *Nature* 477, 203–206.
 - Linnen, C.R., Poh, Y.P., Peterson, B.K., Barrett, R.D., Larson, J.G., Jensen, J.D., and Hoekstra, H.E. (2013). Adaptive evolution of multiple traits through multiple mutations at a single gene. *Science* 339, 1312–1316.
 - Rockman, M.V. (2012). The QTN program and the alleles that matter for evolution: all that's gold does not glitter. *Evolution* 66, 1–17.
 - Steiner, C., Weber, J., and Hoekstra, H. (2007). Adaptive variation in beach mice produced by two interacting pigmentation genes. *PLoS Biol.* 5, e219.
 - Berner, D., Moser, D., Roesti, M., Buescher, H., and Salzburger, W. (2014). Genetic architecture of skeletal evolution in European lake and stream stickleback. *Evolution* 68, 1792–1805.
 - Visscher, P.M. (2008). Sizing up human height variation. *Nat. Genet.* 40, 489–490.

Zoological Institute, University of Basel,
Vesalgasse 1, 4051 Basel, Switzerland.
*E-mail: marius.roesti@unibas.ch

<http://dx.doi.org/10.1016/j.cub.2014.09.005>

Pericentrin: Critical for Spindle Orientation

Mutations in the pericentrin (PCNT) gene cause Majewski osteodysplastic primordial dwarfism type II (MOPDII). Recent work reveals that a discrete set of centrosome proteins require PCNT for their robust localization to mitotic spindle poles. Critically, this complex is crucial for mitotic spindle orientation and involved in the pathogenesis of MOPDII.

Yi Luo¹ and Laurence Pelletier^{1,2,*}

The centrosome is the major microtubule-organizing center of animal cells. It is composed of a centriole pair, which recruits more than 100 proteins, collectively referred to as pericentriolar material (PCM). It participates in the regulation of cell motility, adhesion, intracellular transport and mitotic spindle assembly. During mitosis, centrosomes undergo a dramatic increase in size and nucleation capacity, a process called centrosome maturation. Centrosome maturation potentiates robust mitotic spindle assembly and is a prerequisite for the accurate segregation of chromosomes to progeny cells. Indeed, centrosome and spindle abnormalities are frequently observed in human tumors and are associated with genomic instability.

PCNT is a large, elongated coiled-coil molecule that plays a crucial role in centrosome biogenesis and mitotic spindle assembly [1,2]. PCNT acts as

a scaffold for the recruitment and anchoring of a plethora of PCM proteins including CDK5RAP2, NEDD1 and γ -tubulin ring complexes. Mutations in PCNT are associated with several human disorders including the primordial dwarfism MOPDII [3]. A study published in this issue of *Current Biology* by Chen *et al.* [4] reports a novel role for PCNT in the control of spindle orientation through the recruitment of a specific subset of centrosome components.

Previous genetic linkage analysis revealed that biallelic loss-of-function mutations in PCNT caused MOPDII in all 25 patients [3]. However, the precise molecular mechanisms underlying MOPDII pathology had remained unclear. To address this issue, Chen and colleagues generated PCNT^{-/-} mice and mouse embryonic fibroblasts (MEFs). PCNT^{-/-} mice exhibited known features of MOPDII including small body size, microcephaly, craniofacial developmental anomalies, structural kidney defects and vascular development anomalies. Detailed

analyses of PCNT^{-/-} MEFs and patient-derived epithelial cells revealed a dramatic reduction in the amount of astral microtubules and consequently defects in spindle positioning. Moreover, careful examination of PCNT^{-/-} mice revealed that brain, heart and kidney tissues displayed defects consistent with abnormal asymmetric division and diminished cell proliferation. This phenotype is analogous to microcephaly, where asymmetric divisions produce differentiating cells instead of stem cells, which yields a sharp reduction in the total number of neurons [5].

To provide molecular insights into the spindle positioning defects in PCNT^{-/-} cells, Chen *et al.* surveyed the levels of known centriole and centrosome proteins at spindle poles. Three proteins (CDK5RAP2, Ninein and Centriolin) were most drastically reduced in absence of PCNT. Mutations in Ninein and CDK5RAP2 have been associated with microcephaly, suggesting that these proteins contribute to the MOPDII syndrome though their interplay with PCNT [6,7]. Consistently, the *Drosophila* homologue of CDK5RAP2, Centrosomin (Cnn), is required to maintain mitotic PCM in the vicinity of centriole and to promote astral microtubule formation [8]. Ninein is also required for the maintenance of spindle pole integrity through spatial control of Astrin distribution [9]. In their study, Chen and colleagues show that



Chapter 9

Chapter 9.1

Repeated lake-stream divergence in stickleback life history within a central European lake basin

Repeated Lake-Stream Divergence in Stickleback Life History within a Central European Lake Basin

Dario Moser, Marius Roesti, Daniel Berner*

Zoological Institute, University of Basel, Vesalgasse 1, Basel, Switzerland

Abstract

Life history divergence between populations inhabiting ecologically distinct habitats might be a potent source of reproductive isolation, but has received little attention in the context of speciation. We here test for life history divergence between threespine stickleback inhabiting Lake Constance (Central Europe) and multiple tributary streams. Otolith analysis shows that lake fish generally reproduce at two years of age, while their conspecifics in all streams have shifted to a primarily annual life cycle. This divergence is paralleled by a striking and consistent reduction in body size and fecundity in stream fish relative to lake fish. Stomach content analysis suggests that life history divergence might reflect a genetic or plastic response to pelagic *versus* benthic foraging modes in the lake and the streams. Microsatellite and mitochondrial markers further reveal that life history shifts in the different streams have occurred independently following the colonization by Lake Constance stickleback, and indicate the presence of strong barriers to gene flow across at least some of the lake-stream habitat transitions. Given that body size is known to strongly influence stickleback mating behavior, these barriers might well be related to life history divergence.

Citation: Moser D, Roesti M, Berner D (2012) Repeated Lake-Stream Divergence in Stickleback Life History within a Central European Lake Basin. PLoS ONE 7(12): e50620. doi:10.1371/journal.pone.0050620

Editor: Daniele Canestrelli, Tuscia University, Italy

Received: August 29, 2012; **Accepted:** October 23, 2012; **Published:** December 4, 2012

Copyright: © 2012 Moser et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This work was supported by grants from the Swiss National Science Foundation (Ambizione PZ00P3_126391/1) and the Research Fund of the University of Basel to DB. MR was supported by a Swiss National Science Foundation Sinergia grant (CRSII3_136293) to Walter Salzburger. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: daniel.berner@unibas.ch

Introduction

Speciation is often initiated by adaptation to ecologically distinct habitats in the face of gene flow [1–4]. This process is typically inferred from concurrent divergence in phenotypes and genetic marker frequencies across habitat transitions in the absence of physical dispersal barriers (e.g., [5–13]). Patterns aside, the actual mechanisms constraining gene flow in the early stages of ecological divergence generally remain poorly understood [4,14,15] (but see [16,17]). At least partial reproductive isolation is often assumed to result directly from performance trade-offs associated with adaptive divergence. That is, divergence in ecologically important traits causes selection against maladapted migrants and hybrids between habitats [14,18–20]. Further reductions in gene flow between populations can arise readily as indirect (correlated) consequences of adaptive divergence [4,14,21,22], for instance when traits under ecological divergence also influence reproductive behavior [23–25]. Understanding speciation thus benefits greatly from a thorough understanding of adaptive divergence.

In animals, the traits receiving greatest attention in the context of ecological divergence and reproductive isolation are typically those related to resource acquisition and predator avoidance [14,18]. By contrast, divergence in life history is less frequently considered as a driver of speciation, despite its potential to contribute to reproductive isolation at multiple levels simultaneously: first, adaptive divergence in life history traits in response to ecologically distinct habitats [26,27] might directly reduce gene flow between populations through reduced performance of migrants and hybrids between the habitats. Second, life history

divergence often involves shifts in reproductive timing, thereby potentially causing phenological assortative mating as a correlated response. Evidence of this mechanism exists but is mostly limited to insects (e.g., [28–30]; but see [31]). Third, life history divergence commonly involves body size shifts [26,27]. Because body size is also frequently involved in sexual selection [32], life history divergence might drive sexual assortative mating as an additional correlated response. Finally, life history traits generally display higher levels of phenotypic plasticity than morphological, physiological, and behavioural traits, because the former represent greater targets for environmental perturbation [33,34]. Life history shifts might thus follow rapidly upon the colonization of new habitats, and hence contribute to reproductive isolation well before genetically-based divergence in less plastic traits has occurred [35,36].

The objective of this study is to initiate an investigation of life history divergence in a natural model system for studying speciation with gene flow – lake and stream populations of threespine stickleback fish (*Gasterosteus aculeatus* L.). Marine (ancestral) stickleback have colonized freshwater environments all across the Northern Hemisphere after the last glacial retreat, thereby establishing numerous evolutionarily independent population pairs residing in adjacent lake and stream habitats [37–46]. Lake and stream populations typically display predictable and at least partly genetically-based [39,47,48] divergence in morphological traits, presumably reflecting adaptation to distinct foraging environments. This phenotypic divergence often coincides with striking divergence in genetic markers on a small spatial scale [12,46,49,50], indicating the presence of strong reproductive

barriers associated with lake-stream transitions. The nature of these barriers, however, remains poorly understood (reviewed in [51]).

A contribution of life history divergence to reproductive isolation in lake-stream stickleback, through one or several of the mechanisms described above, is plausible because life history evolution is reported from other stickleback systems. This includes divergence in age at reproduction and reproductive investment within and among lake populations [52–56], and divergence in body size within and among lake populations [52,53,56–59] and between freshwater and marine stickleback [60,61]. At least some of this divergence is partly genetically based [58,62]. Furthermore, body size divergence is generally a strong contributor to mating isolation in the species ([59–61,63–66]; but see [67]. Nevertheless, investigations of life history divergence in lake-stream stickleback are lacking.

Our study focuses on stickleback inhabiting contiguous lake and stream habitats within a single lake basin in Central Europe. We focus on multiple replicate lake-stream sample pairs to assess whether life history divergence has occurred repeatedly in a similar direction. Finally, we include nuclear and mitochondrial genetic marker data to search for signatures of habitat-associated barriers to gene flow, and to gain insight into the origin of lake and stream stickleback populations within the lake basin.

Materials and Methods

Stickleback Samples

The main focus of this life history investigation lies on stickleback in Lake Constance (LC) and its tributaries in Central Europe (Fig. 1, Table 1). The geographic distance between the different lake-stream pairs ('systems') was maximized to reduce the opportunity for gene flow among systems, and to provide phenotypic and genetic information representative of the entire lake basin. The systems include two lake-stream pairs subjected previously to an analysis of foraging morphology and population genetics ('Constance South', COS, and 'Constance West', COW; [44]; see also [68]). The majority of the study sites, however, have not been investigated before. The new systems include 'Constance North' (CON) and 'Constance East' (COE). In the latter, the stream site was sampled at two different locations (Grasbeuren, 7.6 km from the lake, and Mühlhofen, 4 km from the lake). These samples proved very similar phenotypically and genetically (e.g., $F_{ST} = 0.002$, $P = 0.40$; further details not presented), so that they were pooled to represent a single stream site (COE stream). Further, we sampled an additional stream for the COS system ('COS1 stream'). Because this stream drains into LC at almost the same location as COS2 stream, these two systems share their lake counterpart.

The origin of stickleback in the LC basin is unknown, but commonly attributed to human introduction (e.g., [44,69]). The first report of the species' wide-spread occurrence within the basin dates back to the mid 19th century ([70], p. 320). To obtain new genetic insights into the populations' possible origin, we complemented our paired lake-stream samples by samples from two solitary (allopatric) stream-resident populations. The first solitary population was sampled from a small creek draining into the River Rhine (the outlet stream of LC, draining into the Atlantic) near Basel, Switzerland (Fig. 1, Table 1). This sample is hereafter called the Rhine (RH) sample. A recent study indicates strong differentiation in neutral markers between stickleback occurring in the Rhine catchment downstream of LC and the lake itself [69], suggesting that the latter was not colonized via the Rhine. Our Rhine sample allowed an independent evaluation of this hypothesis. The second

solitary stream population (DAN) was sampled in the headwaters of the Danube River drainage near Kirchbierlingen, Germany. This sample was included because of the close proximity of the Danube drainage to the LC basin, and because the LC region drained into the Danube (and eventually into the present-day black sea region) in postglacial times [71].

All new samples were collected in the spring 2011 (late April, May; i.e., during the stickleback breeding season). The samples taken in previous years, and a few specimens collected in 2012 exclusively for the analysis of fecundity and egg size (see below), were also collected within that seasonal time frame. All samples were taken with permission from the corresponding fisheries authorities (Austria: Landesfischereizentrum Vorarlberg, A. Lunardon; Germany: Fischereiforschungsstelle Baden-Württemberg, S. Blank, M. Bopp, C. Wenzel; Switzerland: Jagd- und Fischereiverwaltung Thurgau, R. Kistler; Amt für Umwelt und Energie Basel-Stadt, H.-P. Jermann). Sampling occurred on breeding grounds using unbaited minnow traps. All individuals used for this study were in reproductive stage because the males consistently displayed breeding coloration, and gravid females were frequent at every site. The specimens were euthanized with an overdose of MS-222, taking all efforts to minimize suffering, and immediately weighed, photographed with a reference scale as described in [12], and stored in absolute ethanol. For most sites, a minimum sample of 12 individuals per sex could be achieved (Table 1). Unless noted otherwise, all analyses are based on the full sample from a given site. All work in this study was approved by the Veterinary Office of the Canton of Basel-Stadt (permit number: 2383).

Analysis of Lake-stream Divergence in Life History

Our prime interest was to investigate lake-stream divergence in age and size at reproduction. To quantify age at reproduction, we retrieved the left and right sagittal otolith from all specimens in each lake-stream pair. The otoliths were cleaned mechanically using fine forceps, dried, mounted in 20 μ l Euparal on a microscope slide, and inspected under a stereomicroscope at 50x magnification by a single person (DM) blind to the specimens' origin. Illumination was from above on a black background to optimally visualize the opaque and transparent ring zones used for age determination following [72] (representative otoliths from different age classes are shown in Appendix S1). Left and right otoliths always produced consistent results. A total of 4 specimens (<2% of all specimens investigated) displayed unclear otolith ring patterns and could thus not be aged unambiguously. Excluding these specimens from analysis did not affect any conclusions; hence we present results based on the full data set. Differences in age composition between lake and stream fish were tested separately for each system through non-parametric permutation tests randomizing the response variable (age) 9999 times over the predictor (habitat) [73], and using the lake-stream difference in average age as test statistic. All statistical inference in this study is based on analogous permutation tests.

To quantify body size at reproduction, we digitized 16 homologous landmarks [44] on the photograph of each specimen by using TpsDig [74]. TpsRelw [74] was then used to calculate centroid size from the landmark configurations. This size metric, hereafter referred to as 'body size', was considered more robust to variation in overall body shape and feeding or reproductive status than size metrics such as standard length or linearized body mass. (Using the latter as body size metric, however, produced very similar results in all analyses.) To test for lake-stream divergence in body size, we used the difference in average size between the habitats as test statistic.

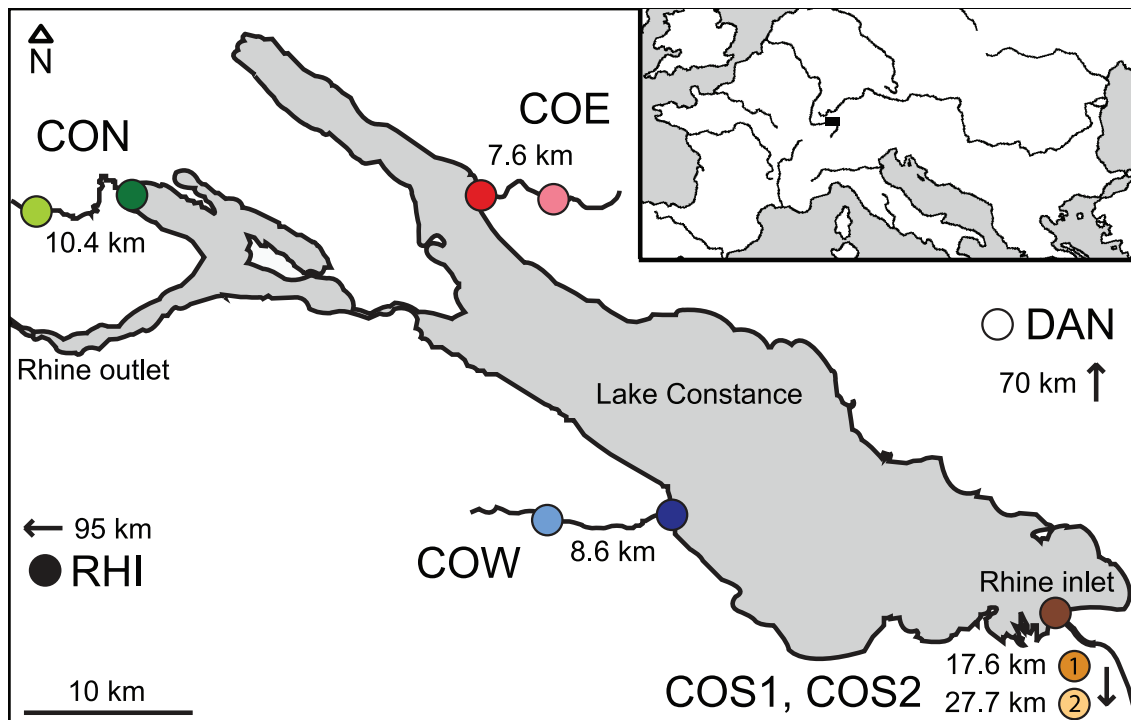


Figure 1. Geographical situation of the stickleback study sites. Shown are the five lake-stream stickleback pairs ('systems') in the Lake Constance basin (CON, COE, COS1, COS2, COW; colored circles, stream sites lighter), and the two solitary sample sites outside the basin (RHI, DAN; black and white circle). The black rectangle in the inset map locates the study area in Central Europe. Distances indicate the approximate water distance between the lake and stream site within each system, and the approximate map distance between Lake Constance and the solitary sample sites. Note that the COS1 and COS2 stream samples were not collected from the Rhine (the major inlet to Lake Constance), but from two small streams draining separately into Lake Constance. Further details on the samples and locations are given in Table 1. doi:10.1371/journal.pone.0050620.g001

Table 1. Localities, geographical coordinates, sampling year, and sample size for the five lake-stream stickleback systems in the Lake Constance basin (CON, COE, COS1, COS2, COW), and the two solitary stream populations (RHI, DAN).

Locality	System or site code	Habitat	Latitude (North)	Longitude (East)	Sampling year	Sample size
Iznang (DE)	CON	lake	47°43'3.36"	8°57'42.48"	2011	22 (10/12)
Bohlingen (DE)	CON	stream	47°43'18.84"	8°53'01.68"	2011	23 (15/7)
Unteruhldingen (DE)	COE	lake	47°43'25.32"	9°13'37.56"	2011	33 (18/15)
Grasbeuren (DE)	COE	stream	47°43'39.72"	9°18'23.4"	2011	13 (9/4)
Mühlhofen (DE)	COE	stream	47°44'11.76"	9°15'49.68"	2011	12 (7/5)
Fussach (AT)	COS1 & COS2	lake	47°29'29.7"	9°39'40.37"	2008	24 (3/21)
Hohenems (AT)	COS1	stream	47°21'18.55"	9°40'10.22"	2008	25 (11/14)
Rankweil (AT)	COS2	stream	47°16'19.28"	9°35'32.72"	2008	24 (12/12)
Romanshorn (CH)	COW	lake	47°33'22.5"	9°22'48.25"	2008/2009	24 (12/12)
Niederaach (CH)	COW	stream	47°33'29.25"	9°16'42.38"	2008/2009	25 (11/14)
Basel (CH)	RHI	stream	47°32'44.34"	7°33'51.84"	2011	24 (12/12)
Kirchbierlingen (DE)	DAN	stream	48°14'04.03"	9°43'30.86"	2011	34 (15/19)

The localities are situated in Germany (DE), Austria (AT), and Switzerland (CH). Sample sizes are total, and males and females in parentheses. Note that the same lake sample was used for both the COS1 and COS2 system, and that the COE stream site combines two samples (for details see text). doi:10.1371/journal.pone.0050620.t001

In addition to age and size at reproduction, we investigated divergence in fecundity and egg size. For this, clutches of gravid females ready for spawning were collected in the field by gently squeezing the females' abdomen, and preserved in ethanol. We then counted the total number of eggs (fecundity) under a stereomicroscope, dried all eggs at 50°C for 48 h, and determined their total dry mass. Egg size was then expressed as the total clutch dry mass divided by total egg number (i.e., the average dry mass of a single egg). This investigation used mainly females collected in 2012 for this specific purpose only (and hence not included in Table 1; lake: COE, COW, N = 11 each; stream: COW, CON, COE, N = 9, 1, 1), but additionally involved a few females also used for the other analyses (details given in Table S1). Testing for lake-stream divergence in fecundity and egg size was then performed in a single analysis for each trait by pooling data across the two lake sites and the three stream sites. (Restricting the analysis to the COW system with sufficient data from each habitat produced similar results.) As above, the difference in trait means between the habitats was used as test statistic.

Comparison of Body Size Among Global Populations

To interpret the body size patterns revealed in our lake-stream and solitary stickleback populations from Central Europe in a broader geographic and ecological context, we performed a comparison of reproductive body size by including a total of 21 additional stickleback populations from different geographic regions and habitats. We hereafter call this the 'global' data set, acknowledging that these samples do not represent the species' full body size diversity (e.g., [52]). These additional samples comprised lake populations from Beaver, Boot, Joe's, Misty, Morton, Pye, and Robert's Lake (sites described in [43]), and from Hope Lake (coordinates: 50°34'0" N, 127°20'30" W), on Vancouver Island (British Columbia, Canada). Additional stream-resident populations were from the Beaver, Boot, Joe's, McCreight, Pye, and Robert's systems [43], and from the inlet stream to Misty Lake [39,75], on Vancouver Island. These freshwater samples were complemented by collections of marine stickleback from two estuaries on the east coast of Vancouver Island (Cluxewe: 50°36'51" N, 127°11'10" W; Sayward [76]), from the Japan Sea and Pacific [77], from the Atlantic Coast in Norway [78], and from the coast of the White Sea in Russia [79]. All these additional samples were also collected during the reproductive season on breeding grounds. Body size was quantified from available photographs as described above. Sample size was 20–36 individuals per site, with both sexes well represented.

For the global comparison of body size at reproduction, we first pooled all samples from the LC basin within each habitat type. This was done to avoid pseudo-replication, and because body size within each habitat type was highly consistent (see below). Interestingly, visual inspection of the data from the global samples suggested differences among the three habitat types (lake, stream, marine) in the *variability* of average body size across populations. This was tested formally through separate lake-stream and marine-stream tests using the variance in population means as test statistic.

Additional Phenotypic Analyses

The above analyses were complemented by investigating two additional variables potentially relevant to life history evolution. First, as life history divergence might be driven by differential food resources, we analyzed prey items in stomachs of stickleback from one system (COW lake and stream; N = 20 and 7). Because lake stickleback might exploit different prey resources during the reproductive period spent in littoral (near-shore) breeding habitat than during non-reproductive life stages (e.g., [80]), we addition-

ally acquired a small sample (N = 5) of stickleback caught by LC fishermen in offshore drift nets targeting pelagic whitefish. This sample was taken off the COS lake site in April 2011. To ensure adequate quality of stomach content for analysis, all specimens (lake offshore, lake littoral, and stream) were preserved within 5 h upon setting the capturing device (minnow trap, drift net). Prey items were identified to order, family, or genus, and assigned to broad taxonomic groups (e.g., pelagic cladocera, vermiform insect larvae; see Table 2). For every stickleback, we determined the relative proportion of the total prey items accounted for by each taxonomic group, calculated summary statistics for each of the three habitat types, and interpreted these statistics qualitatively. This approach was preferred to a formal analysis because of the relatively small sample sizes.

The second additional variable was the lateral plate phenotype. Ancestral marine stickleback are protected from vertebrate predators in their pelagic environment by bony lateral plates along their entire body [81]. This phenotype is disfavoured in most freshwater environments, as stickleback in lakes and streams generally display an adaptive, genetically-based reduction in the number of lateral plates [81]. We considered this trait here because the major genetic factor determining plate phenotype (the ectodysplasin gene, *EDA*; [82]) might pleiotropically influence growth rate [83], and because stickleback in the LC basin are polymorphic for both plate phenotype and the underlying *EDA* alleles [44]. Following this latter study, we assigned all individuals to one of three lateral plate phenotype morphs (full, partial, low). We then tested for lake-stream divergence in plate morph frequency within each system by using the Chi-square ratio as test statistic (extending similar tests already performed for the COW and one of the COS systems; [44]). Next, sufficiently polymorphic samples (i.e., the stream samples of CON, COE, and COW) were used to test for an association between plate morph and body size by using the F ratio from analysis of variance as test statistic [73]. All statistical analyses and plotting were performed in R ([84]; codes available on request). All phenotypic data are provided in Table S1.

Genetics

The major goal of our genetic investigation based on nuclear and mitochondrial markers was to quantify population structure within and among the replicate lake-stream systems in the LC basin. Of particular interest was the detection of strong genetic divergence within lake-stream systems, suggesting effective habitat-related barriers to gene flow. An additional goal was to explore the relationship between stickleback in the LC basin and fish from nearby water bodies. The present work greatly extends a previous population genetic study partly involving fish from the LC basin [44] in that new lake-stream pairs are analyzed, samples from the Rhine and Danube are included, and a greater number of genetic markers are used.

We first extracted DNA from pectoral and caudal fin tissue on a MagNA Pure LC extraction robot (Roche) by using the Isolation Kit II (tissue). Next, we amplified eight microsatellites with labelled primers in two separate multiplex PCRs by using the QIAGEN multiplex kit and following the manufacturer's protocol. All PCRs included a negative control to check for contamination. The microsatellite markers were chosen to be far from known quantitative trait loci in stickleback, and to lie on different chromosomes. They included the markers Stn67, Stn159, Stn171, and Stn195 used previously [12,44], and additionally Stn28, Stn99, Stn119, and Stn200 [85]. For the latter, we designed our own primer pairs (primer sequences for all eight markers are provided in Table S2). PCR products were run on an ABI3130xl

Table 2. Stomach content of stickleback from the Lake Constance offshore site, and from the lake and stream site in the COW system.

	Pelagic	Pelagic or benthic	Benthic				
	Cladocera ¹	Copepods	Cladocera ²	Other crustacea ³	Vermiform insect larvae ⁴	Other insect larvae ⁵	Stickleback eggs
Lake offshore	0.34 (0.21)	0.66 (0.21)	–	–	–	–	–
COW lake	0.01 (0.02)	0.07 (0.1)	0.33 (0.29)	0.03 (0.08)	0.42 (0.37)	0.15 (0.24)	0.03 (0.11)
COW stream	–	0.17 (0.18)	0.2 (0.25)	–	0.57 (0.27)	0.06 (0.08)	0.09 (0.2)

¹Daphnia, Ceriodaphnia, Bosmina.

²Chydoridae.

³mainly Ostracoda.

⁴Chironomidae, Ceratopogonidae.

⁵mainly Ephemeroptera and Plecoptera.

The values represent the proportion of the total prey items accounted for by each prey class, averaged across individuals within each site (standard deviation in parentheses). The copepods category subsumes pelagic, benthic, and/or generalist taxa difficult to distinguish; strictly pelagic calanoid copepods, however, were found in the offshore lake specimens only. Sample size is 5, 20, and 7 for offshore, COW lake, and COW stream.

doi:10.1371/journal.pone.0050620.t002

sequencer (Applied Biosystems), and alleles scored manually in PeakScanner v1.0. Input files for the different population genetic programs were prepared by using CREATE [86].

The microsatellite data were first used to estimate differentiation among all 11 samples by Weir & Cockerham's F_{ST} [87] calculated with GENETIX v4.0.5.2 [88] (P-values based on 999 permutations). To account for variation in heterozygosity within populations [89], we also calculated *standardized* F_{ST} after data transformation with RECODEDATA v0.1 [90]. Next, we tested whether neighboring lake and stream samples qualified as genetically distinct populations by performing a genetic clustering analysis using STRUCTURE (v2.3.1; [91,92]) separately in each lake-stream pair (note that the COS system represents two pairs, both involving the same lake sample). The assumed number of populations (K) ranged from one to three, with each level replicated five times under the admixture and independent allele model with 100'000 iterations (20'000 iterations burnin). An additional analysis examined population structure among the 11 pooled samples, using $K = 1-12$. STRUCTURE results were combined using Structure Harvester v.0.6.92 [93], and interpreted following [94,95]. The microsatellite data set is provided in Table S3.

The above analyses using rapidly evolving microsatellites were complemented by a more coarse-grained investigation of genetic relationships based on single nucleotide polymorphisms (SNPs) within a 305 bp segment of the mitochondrial D-loop. Sample size was 18–32 individuals per site, 256 in total. Primers and PCR amplification conditions were as in [44]. Products were sequenced on an ABI3130xl sequencer (Applied Biosystems). We used jModelTest v0.1.1 [96] to determine the most appropriate model of sequence evolution ('F81'; [97]), identified the most probable genealogical relationship by the maximum-likelihood method implemented in PAUP* v4.0 [98], and generated a haplotype genealogy for visualization following [99]. All D-loop sequences are deposited in GenBank (accession numbers JX436521–JX436776).

Results

Phenotypic Analyses

The otolith analysis revealed strong and highly consistent lake-stream divergence in age at reproduction in all replicate systems in the LC basin (all $P < 0.0015$). Generally, stickleback on breeding

grounds in the lake were in their third calendar year (i.e., approximately two years old), with a few individuals breeding in their second or fourth calendar year (Fig. 2). By contrast, stream stickleback essentially displayed an annual life cycle; individuals in their third calendar year were rare, and no single fish was found to breed in its fourth calendar year.

Lake-stream shifts in age at reproduction were paralleled by strong divergence in body size, with lake fish on average exhibiting 27% greater size than stream fish (lake mean centroid size across all systems: 80.4 mm; stream: 63.2 mm; $P = 0.0001$ in all systems) (Fig. 2). Translated to fresh body mass, the average size difference was more than twofold (lake: 2.53 g; stream 1.19 g; a photograph of a representative lake and stream individual is shown in Appendix S1). Body size divergence was further associated with dramatic divergence in fecundity (Fig. 3); on average, the (larger) lake females displayed a threefold higher number of eggs than the stream females (284 *versus* 94; $P = 0.0001$). Egg size, however, did not differ between the habitats ($P = 0.51$).

Our comparison of body size across global stickleback samples from lakes, streams, and the sea indicated a clear difference in the variance in population average size among the habitats. Strikingly, all stream populations investigated displayed relatively similar average size, whereas the lake samples were much more variable (lake-stream difference in variance: $P = 0.002$; Fig. 4). The latter included very small-bodied populations (Morton, Pye, and Robert's) as well as large-bodied populations (Boot, Joe's). Body size among marine stickleback also tended to be more variable than among stream populations (marine-stream difference in variance: $P = 0.065$; note the small sample size for marine fish, and hence low statistical power in this test).

In addition to the above life history patterns, our analysis of stomach content revealed a very clear difference in prey utilization by lake and stream stickleback, despite the modest sample sizes. In particular, our pelagic sample showed clearly that LC stickleback forage on zooplankton outside the breeding grounds; the stomachs of these specimens contained exclusively small pelagic crustacea (Table 2). By contrast, the stomachs of the stream fish contained exclusively benthic prey (predominantly chironomid larvae and benthic cladocera), highly consistent with data from streams on Vancouver Island [43]. Similar benthic prey was also found in the lake fish collected on (littoral) breeding grounds, indicating a reproductive shift in foraging mode in stickleback residing within LC.

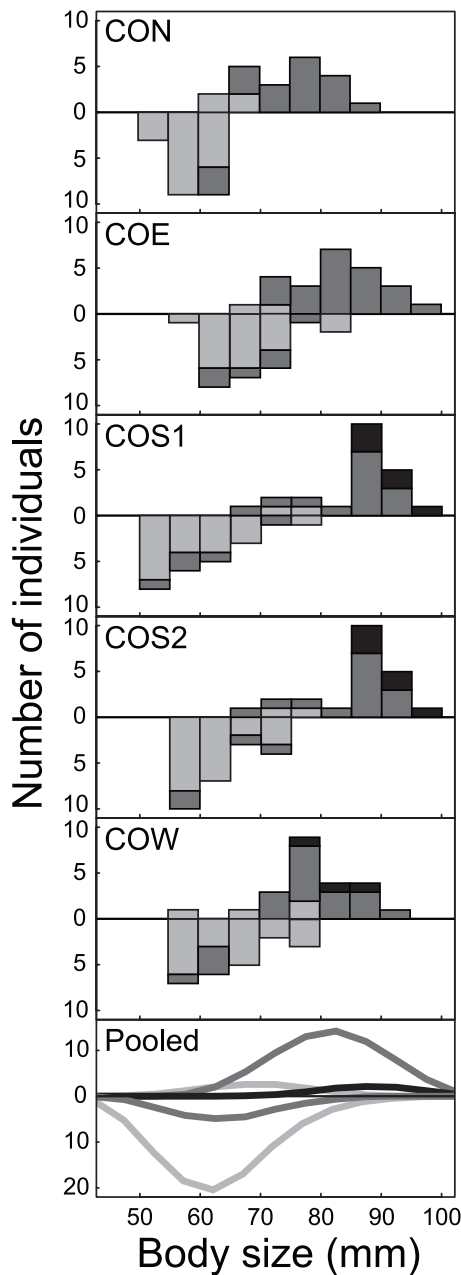


Figure 2. Age and body size at reproduction in lake and stream stickleback from the Lake Constance basin. The top panels show body size (quantified as landmark-based centroid size) histograms for each lake-stream system separately, with the lake data pointing upward and the stream data pointing downward. Proportions are shaded according to age class; individuals in their second, third, and fourth calendar year are drawn in light gray, dark gray, and black. The bottom panel follows the same drawing conventions, except that here the data are pooled across all systems within each habitat type, and smoothed

by LOESS (locally weighted scatterplot smoothing) for each age class separately. Note the striking shift toward greater age and size at reproduction in lake stickleback as compared to their conspecifics from streams.

doi:10.1371/journal.pone.0050620.g002

In all three new lake-stream systems subjected to lateral plate morph analysis (CON, COE, COS1), we found a trend toward plate reduction in the stream as compared to the lake where fully plated fish predominated clearly. The shift in plate morph frequency was particularly striking in the COE system ($P = 0.0001$), paralleling a similar pattern found previously in the COW system [44] (details given in Appendix S1). However, we found no relationship between plate morph and body size at reproduction in any of the three investigated stream samples (CON, COE, COW; all $P > 0.35$).

Genetics

A striking pattern revealed by our eight microsatellite markers was the absence of population structure among the four geographically distant LC samples. None of the six total pairwise F_{ST} values among these lake samples exceeded 0.01 (all $P > 0.07$) (Table 3). Genetic differentiation *within* the lake-stream pairs was mostly modest as well, but sometimes reached substantial values despite a much shorter geographic distance between the paired lake and stream sites than among the lake sites (COE: $F_{ST} = 0.18$, $P = 0.001$; COS2: $F_{ST} = 0.08$, $P = 0.001$). Microsatellite differentiation among the stream samples was generally substantial, with F_{ST} averaging 0.10 (all $P < 0.004$ except CON-COS1, $P = 0.13$). Furthermore, our Rhine sample (RHI) displayed strong differentiation from all samples in the LC basin ($F_{ST} = 0.16-0.29$), whereas differentiation between the Danube sample (DAN) and stickleback from the LC basin was rather low. For instance, all five comparisons between DAN and LC samples produced $F_{ST} < 0.04$ ($P = 0.001-0.023$).

The results from the STRUCTURE analysis agreed well with the F_{ST} -based patterns. First, analyzing each system separately, STRUCTURE identified the system displaying the highest lake-stream differentiation (COE) as consisting of two genetically distinct populations. The four other systems qualified as a single population (details not presented). Analyzing all 11 samples together suggested two distinct genetic clusters. The first cluster involved RHI and the stream site of COE, the second involved all other populations from the LC basin plus the DAN sample.

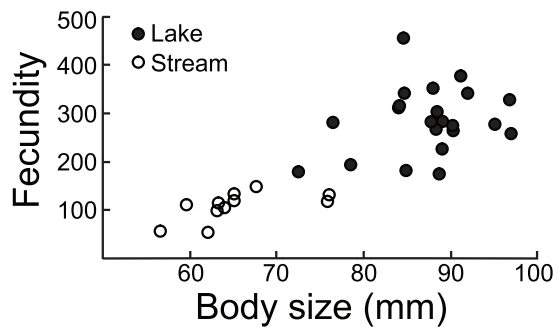


Figure 3. Fecundity in relation to body size in female stickleback from Lake Constance and its tributary streams. Fecundity is expressed as number of eggs per clutch. Within each habitat class, samples were pooled across different locations (lake: $N = 22$; stream: $N = 11$). doi:10.1371/journal.pone.0050620.g003

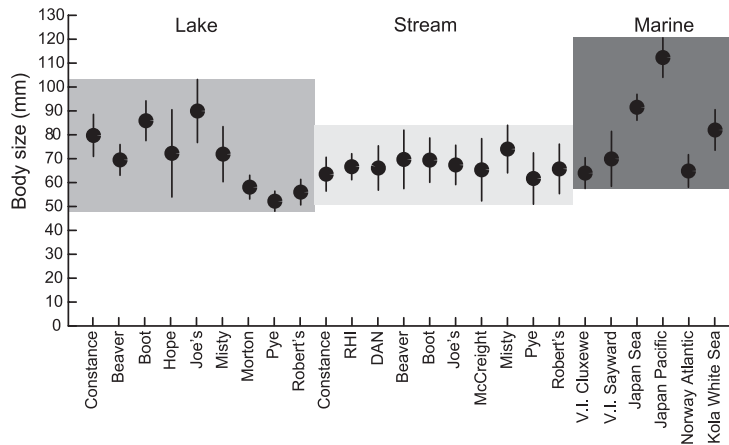


Figure 4. Body size at reproduction in the global stickleback populations from lake, stream, and marine habitats. Samples from the Lake Constance basin are pooled for each habitat type (further details on the samples are given in the text). Error bars are one standard deviation in each direction. The shaded boxes behind the symbols indicate the body size range spanned by the standard deviations in each habitat. Note the low variance in population mean size among the stream populations as compared to lake and marine fish. doi:10.1371/journal.pone.0050620.g004

However, the STRUCTURE algorithm can perform poorly when faced with highly imbalanced sample sizes [100]. Indeed, most samples from the LC basin were genetically so similar that they essentially formed one single large sample, which probably caused RHI and COE stream to cluster together despite strong genetic differentiation ($F_{ST} = 0.16$). However, when analyzing only RHI, COE stream, and a *single* lake sample together, three distinct populations were indicated, as expected based on F_{ST} .

Our mitochondrial D-loop sequencing identified six total SNPs, defining five distinct haplotypes (Fig. 5). One of these haplotypes was clearly predominant; it was either the only one discovered, or at least very frequent, in *all* samples from the LC basin. Notably, this haplotype was also the only one found in the DAN sample. By contrast, all individuals from RHI exhibited a different haplotype

shared only with some individuals from three stream samples of the LC basin. Three additional haplotypes occurred at low frequency, mainly in stream fish.

Discussion

Life History Divergence and Implications for Reproductive Isolation

Divergence in life history traits might strongly contribute to reproductive isolation, and yet its role in speciation is little explored. We here investigated life history in stickleback residing in Lake Constance and multiple tributary streams, revealing dramatic divergence between the two habitats: lake fish reproduce at much greater age and size than their conspecifics in the streams,

Table 3. Pairwise genetic differentiation among the nine lake and stream stickleback samples from the Lake Constance basin, and the two solitary samples, based on eight microsatellite markers.

	CON lake	CON stream	COE lake	COE stream	COS lake	COS1 stream	COS2 stream	COW lake	COW stream	RHI	DAN
CON lake		0.00 (0.676)	0.01 (0.071)	0.18 (0.001)	0.01 (0.240)	0.02 (0.041)	0.10 (0.001)	0.00 (0.305)	0.05 (0.001)	0.27 (0.001)	0.03 (0.002)
CON stream	0.00		0.00 (0.587)	0.15 (0.001)	0.00 (0.386)	0.01 (0.132)	0.06 (0.001)	0.00 (0.759)	0.03 (0.004)	0.25 (0.001)	0.02 (0.011)
COE lake	0.02	0.00		0.18 (0.001)	0.00 (0.543)	0.02 (0.003)	0.07 (0.001)	0.00 (0.744)	0.04 (0.001)	0.28 (0.001)	0.03 (0.001)
COE stream	0.55	0.46	0.50		0.20 (0.001)	0.17 (0.001)	0.21 (0.001)	0.17 (0.001)	0.13 (0.001)	0.16 (0.001)	0.17 (0.001)
COS lake	0.02	0.00	0.00	0.56		0.01 (0.160)	0.08 (0.001)	0.00 (0.478)	0.03 (0.001)	0.28 (0.001)	0.04 (0.001)
COS1 stream	0.05	0.02	0.05	0.47	0.02		0.06 (0.001)	0.02 (0.053)	0.03 (0.002)	0.24 (0.001)	0.08 (0.001)
COS2 stream	0.22	0.13	0.15	0.52	0.17	0.12		0.08 (0.001)	0.11 (0.001)	0.29 (0.001)	0.12 (0.001)
COW lake	0.00	0.00	0.00	0.48	0.00	0.05	0.17		0.02 (0.007)	0.26 (0.001)	0.02 (0.023)
COW stream	0.13	0.08	0.10	0.40	0.07	0.07	0.25	0.05		0.21 (0.001)	0.06 (0.001)
RHI	0.69	0.64	0.66	0.46	0.66	0.56	0.62	0.62	0.54		0.26 (0.001)
DAN	0.08	0.05	0.07	0.50	0.00	0.19	0.27	0.05	0.16	0.65	

The upper semimatrix gives Weir & Cockerham's F_{ST} estimator [87], with P-values based on 999 permutations in parentheses (bold if $P < 0.01$). The lower semimatrix presents F_{ST} standardized by the maximum differentiation possible given the observed magnitudes of within-population heterozygosity [89]. doi:10.1371/journal.pone.0050620.t003

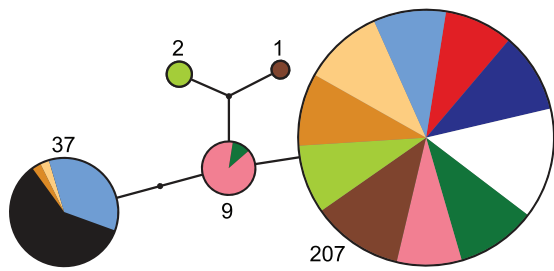


Figure 5. Haplotype network for the lake-stream stickleback pairs in the Lake Constance basin and the solitary populations. The network is based on six single nucleotide polymorphisms in the mitochondrial D-loop. The numbers give the total count for each haplotype. Color codes are as in Fig. 1. doi:10.1371/journal.pone.0050620.g005

and these patterns coincide with much greater fecundity in females from the lake. These findings parallel concurrent shifts in age and size at reproduction and in reproductive investment reported from North American lake populations [52,55,56]. The only life history trait that proved stable between lake and stream stickleback was egg size, possibly indicating similar stabilizing offspring viability selection in both habitats [101,102].

Divergence in age and size at reproduction was highly consistent across multiple replicate habitat pairs in the LC basin, and our genetic data indicate clearly that this results from repeated evolution in stream stickleback. The reason is that the stream samples consistently displayed strong mutual microsatellite differentiation, contrary to the lake samples exhibiting negligible differentiation. This pattern clearly rules out the possibility that the different stream populations originate from a common ancestral stream stickleback population. Moreover, the rare D-loop haplotypes found in the LC basin were mostly unique to specific stream samples (Fig. 5), consistent with independent founder events (i.e., haplotype frequency shifts caused by strong genetic drift in the small stream founder populations). Together, our life history and genetic data thus argue strongly for the independent colonization of the different tributaries by an essentially panmictic LC population, followed by repeated life history evolution in stream stickleback.

Given the great magnitude of lake-stream divergence in body size, and the general importance of this trait in mate choice and male aggressive interactions in the species [59–61,63–66], the observed life history shifts might well contribute to reducing gene flow across the lake-stream habitat transitions. Indeed, our F_{ST} -based analysis revealed substantial lake-stream differentiation within some systems (with values reaching 0.18), and STRUCTURE identified two distinct populations in one of them. This allows us to infer the presence of strong reproductive barriers at a small spatial scale, consistent with findings from lake-stream systems in Pacific North America [12,46,49,50]. Note that the *weak* marker divergence seen in some of our systems (CON, COS1; $F_{ST} < 0.01$) does not conflict with this conclusion; because the colonization of the LC basin is presumably relatively recent (see below), detecting reproductive isolation with neutral markers is expected to be difficult [44,103]. The presence of effective habitat-related reproductive barriers is also supported by the consistent and sometimes substantial (COE, COW) lake-stream divergence in plate morph frequency (Appendix S1). This divergence has a strong genetic basis [44] and would not have arisen, or be maintained, in the absence of effective barriers to gene flow.

Nevertheless, the extent to which the observed lake-stream shifts in life history actually contribute to reproductive isolation cannot be evaluated based on the present data.

Mechanisms of Life History Divergence

In many organisms, the transition of resource allocation from growth to reproductive life is governed by critical maturation size thresholds (reviewed in [104,105]). Although not investigated in detail, this seems to hold for stickleback as well [106,107]: as long as an individual has not attained this threshold, environmental cues signalling spring conditions will not trigger maturation and reproductive behavior. On the basis of this maturation control, we propose two not mutually exclusive hypotheses explaining life history divergence in lake-stream stickleback in the LC basin. First, assuming similar growth rates in both habitats, lake fish might exhibit a relatively higher maturation size threshold (due to genetic divergence and/or phenotypic plasticity) that they generally cannot attain within one year. Only after two years of growth, lake fish would exceed their maturation threshold and start reproducing – and at that time also be much larger than the stream fish reaching their threshold size within one year [105]. This hypothesis is plausible: body size divergence among populations of *minespine* stickleback is attributable to genetically-based divergence in maturation size thresholds [108,109].

Alternatively, maturation size thresholds might be similar among the populations, but growth rates might be lower in lake fish than in tributary stream populations (again due to genetic divergence, phenotypic plasticity, or both). The consequence would be the same as above: lake fish would require two years of growth to attain their maturation threshold, but mature larger [105]. Indeed, our study provides evidence of differential growth rates between the habitats. As the analysis of stomach content suggests, stickleback inhabiting LC exploit exclusively zooplankton prey outside the breeding grounds. These fish are also an occasional by-catch in off-shore drift nets (personal communications from LC fishermen), and are absent from littoral habitat outside the breeding season (D. Moser, personal observation). Moreover, for a freshwater population, stickleback in LC display extremely long gill rakers [44], a character state generally associated with zooplankton exploitation [110] and typical of pelagic marine stickleback [76]. Stickleback residing within LC thus display a pelagic life style, with a foraging niche shift during the reproductive period (see also [80]). Note also that the LC fish provide a rare example of a freshwater population almost fixed for the full lateral plate morph (Appendix S1), a phenotype presumably favored in pelagic populations highly exposed to vertebrate predation [111]. (We found no evidence, however, for a direct relationship between plate phenotype and life history traits.)

By contrast, stream populations in the LC basin exploit exclusively benthic resources. Within the LC basin, we thus find similarly strong divergence in foraging modes as seen in the most ecologically divergent lake-stream pairs on Vancouver Island, Canada [12,43,49]. This difference in resource use might directly induce differential growth performance between the habitats, as benthic foraging generally seems to allow for a higher growth rate than pelagic foraging [112,113]. Direct evidence for divergence in growth rates comes from a small sample of juvenile stickleback captured during the breeding season at the edge of the breeding ground at the COE lake site (non-reproductive status was confirmed by dissection; testes and ovaries were poorly developed). These fish displayed body sizes clearly below those of stream stickleback (43–49 mm, $N = 3$), and yet otolith analysis confirmed that they were already one year old (data presented in Table S1). It

thus appears plausible that a lower growth rate in lake stickleback, induced by a relatively poor pelagic resource base, underlies the lake-stream divergence in life history observed within the LC basin (acknowledging the possibility that differential growth rates in the two habitats has a genetic component).

The direct induction or genetically based evolution of an annual life cycle in response to more profitable benthic resources in streams would explain the relatively low variance in average body size across stream populations from different geographic regions (Fig. 4). The reason is that the resource spectrum used by stream stickleback is highly consistent across global populations, while lake populations are more variable in resource use [12,43,49,114]. If variation in population mean size was (at least partly) a consequence of resource-dependent variation in growth rate, we would indeed expect lake population means to be more variable than stream means. We note, however, that *small-sized* lake populations are not necessarily benthic-foraging. For instance, the lake population with the smallest average size in Fig. 4 (Pye Lake, Vancouver Island) exploits a strictly pelagic food base [43]. Hence, factors other than food resources (e.g., predation [57,58]) likely contribute to the presumably greater life history diversity in lake (and perhaps marine) stickleback than in stream stickleback.

Body size divergence through resource-mediated plasticity in growth rate might play a particularly important role in reproductive isolation. The reason is that this divergence would occur, and potentially influence sexual interactions, within a *single* generation after the colonization of a stream by lake fish [35,36]. It would therefore be crucial to quantify environmental and genetic contributions to life history divergence in stickleback from the LC basin and elsewhere.

Origin of Stickleback in the Lake Constance Basin

Consistent with a previous population genetic investigation [69], our genetic analyses indicate that the populations in the LC basin do not originate from colonization by stickleback residing in the Rhine downstream of LC. However, we find that stickleback in the LC basin are genetically very closely related to those occurring in the nearby Danube drainage: pairwise differentiation between Lake Constance samples and DAN was consistently low ($F_{ST} < 0.04$), and the only D-loop haplotype found in DAN was the one also predominant in the LC basin. Is it possible that LC stickleback derive from a source population from the Black Sea region that colonized naturally via the Danube? A population genetic study in European perch (*Perca fluviatilis*) [71] and geological data [115] suggest the existence of such a temporary colonization route during the last glacial retreat. In fact, a connection between the Danube drainage and the LC basin still persists today, as the source of the stream sampled at the CON stream site is formed by water captured from the Danube headwaters through a sinkhole and a 12 km underground stream [116]. Whether this allows for fish dispersal has not been investigated.

A scenario of colonization via the Danube, however, is challenged by the absence of stickleback from the entire Danube drainage reported in the nineteenth century ([70], p. 319; the species was already present in the LC basin at that time), although the reliability of this information is unknown. Moreover, stream-resident stickleback are generally low-plated (e.g., [38,42,117–119]). The incomplete shifts toward the low-plated morph in our stream samples from the LC basin, along with the low haplotype diversity within the basin, might thus be taken as tentative support of a relatively recent origin, perhaps due to human introduction.

More extensive phylogeographic data from Central and Eastern European populations are needed for a better understanding of the origin and age of stickleback in the LC basin and the Danube drainage.

Conclusions

We have shown strong, repeated, and possibly rapid life history divergence between lake and stream stickleback in the Lake Constance basin, sometimes coinciding with substantial differentiation in neutral markers. Our comparison of body size patterns across global populations and habitats, combined with data from other stickleback systems, further suggests that life history divergence is very common in this species. Our study opens up several important avenues for further investigation: first, experimental work should uncover the mechanistic basis of life history shifts; are they due to differences in maturation size thresholds, in growth rate, or both? Second, the relative contribution of phenotypic plasticity *versus* genetic change to life history divergence should be quantified, and the ecological basis of divergence (e.g., contrasting trophic environments, differential predation regimes) should be identified. Finally, great efforts will be needed to understand whether life history divergence is primarily an aspect of adaptive divergence facilitated by already existing barriers to gene flow, or whether life history divergence itself is a major source of reproductive isolation between lake and stream populations.

Supporting Information

Table S1 Sample site information and phenotypic data for all individuals.
(TXT)

Table S2 Primer sequences used to amplify the eight microsatellite markers.
(TXT)

Table S3 Microsatellite allele data for all individuals included in the population genetic analysis.
(TXT)

Appendix S1 Contains figures displaying representative stickleback otoliths of different ages, illustrating lake-stream divergence in body size, and summarizing lateral plate morph data for all study sites.
(PDF)

Acknowledgments

This work benefited greatly from many people who helped sample and/or provided access to the study populations: H. Bandel, S. Blank, J. Behrmann-Godel, M. Bopp, G. Bosshart, A.-C. Grandchamp, H. Höchstädter, M. Hohler, H.-P. Jermann, R. Kistler, M. Konrad, F. Ley, A. Lunardon, and C. Wenzel. Photographs for the global analysis were contributed by D. Bolnick, A. Hendry, J. Kitano, T. Leinonen, J.-S. Moore, and J. Merilä. K. Lucek, D. Kaiser, and N. Strebel shared their experience in otolith analysis. W. Salzburger provided wet lab resources and infrastructure, and B. Aeschbach and N. Boileau facilitated wet lab work. M. Matschiner created the haplotype network with his unpublished software. W. Salzburger and L. Schärer provided valuable suggestions on the manuscript. We are most grateful for all this support.

Author Contributions

Conceived and designed the experiments: DB DM MR. Performed the experiments: DM DB MR. Analyzed the data: DM DB. Wrote the paper: DB DM.

References

1. Endler JA (1977) Geographic variation, speciation, and clines. Princeton: Princeton University.
2. Gavrillets S, Li H, Vose MD (2000) Patterns of parapatric speciation. *Evolution* 54: 1126–1134.
3. Wu CI (2001) The genic view of the process of speciation. *J Evol Biol* 14: 851–865.
4. Sobel JM, Chen GF, Watt LR, Schemske DW (2010) The biology of speciation. *Evolution* 64: 295–315.
5. Smith TB, Wayne RK, Girman DJ, Bruford MW (1997) A role for ecotones in generating rainforest biodiversity. *Science* 276: 1855–1857.
6. Lu G, Bernatchez L (1999) Correlated trophic specialization and genetic divergence in sympatric lake whitefish ecotypes (*Coregonus clupeaformis*): support for the ecological speciation hypothesis. *Evolution* 53: 1491–1505.
7. Ogden R, Thorpe RS (2002) Molecular evidence for ecological speciation in tropical habitats. *Proc Natl Acad Sci USA* 99: 13612–13615.
8. Barluenga M, Stoltzing KN, Salzburger W, Muschick M, Meyer A (2006) Sympatric speciation in Nicaraguan crater lake cichlid fish. *Nature* 439: 719–723.
9. Grahame JW, Wilding CS, Butlin RK (2006) Adaptation to a steep environmental gradient and an associated barrier to gene exchange in *Littorina saxatilis*. *Evolution* 60: 268–278.
10. Foster SA, McKinnon GE, Steane DA, Potts BM, Vaillancourt RE (2007) Parallel evolution of dwarf ecotypes in the forest tree *Eucalyptus globulus*. *New Phytol* 175: 370–380.
11. Seehausen O, Terai Y, Magalhaes IS, Carleton KL, Mrosso HDJ, et al. (2008) Speciation through sensory drive in cichlid fish. *Nature* 455: 620–U623.
12. Berner D, Grandchamp A-C, Hendry AP (2009) Variable progress toward ecological speciation in parapatry: stickleback across eight lake-stream transitions. *Evolution* 63: 1740–1753.
13. Rosenblum EB, Harmon LJ (2011) “Same same but different”: replicated ecological speciation at White Sands. *Evolution* 65: 946–960.
14. Coyne JA, Orr HA (2004) Speciation. Sunderland: Sinauer Associates.
15. Rundle HD, Nosil P (2005) Ecological speciation. *Ecol Lett* 8: 336–352.
16. Ramsey J, Bradshaw HD, Schemske DW (2003) Components of reproductive isolation between the monkeyflowers *Mimulus lewisii* and *M. cardinalis* (Phrymaceae). *Evolution* 57: 1520–1534.
17. Nosil P (2007) Divergent host plant adaptation and reproductive isolation between ecotypes of *Timema cristinae* walking sticks. *Am Nat* 169: 151–162.
18. Schluter D (2000) The ecology of adaptive radiation. Oxford: Oxford University.
19. Hendry AP (2004) Selection against migrants contributes to the rapid evolution of ecologically dependent reproductive isolation. *Evol Ecol Res* 6: 1219–1236.
20. Nosil P, Vines TH, Funk DJ (2005) Reproductive isolation caused by natural selection against immigrants from divergent habitats. *Evolution* 59: 705–719.
21. Rice WR, Hostert EE (1993) Laboratory experiments on speciation: what have we learned in 40 years? *Evolution* 47: 1637–1653.
22. Gavrillets S (2004) Fitness landscapes and the origin of species. Princeton: Princeton University.
23. Ritchie MG (2007) Sexual selection and speciation. *Annu Rev Ecol Evol Syst* 38: 79–102.
24. Bonduriansky R (2011) Sexual selection and conflict as engines of ecological diversification. *Am Nat* 178: 729–745.
25. Maan ME, Seehausen O (2011) Ecology, sexual selection and speciation. *Ecol Lett* 14: 591–602.
26. Stearns SC (1992) The evolution of life histories. Oxford: Oxford University Press.
27. Roff DA (2002) Life history evolution. Sunderland: Sinauer.
28. Feder JL, Roethele JB, Wlazlo B, Berlocher SH (1997) Selective maintenance of allozyme differences among sympatric host races of the apple maggot fly. *Proc Natl Acad Sci USA* 94: 11417–11421.
29. Abbot P, Withgott JH (2004) Phylogenetic and molecular evidence for allochronic speciation in gall-forming aphids (*Pemphigus*). *Evolution* 58: 539–553.
30. Santos H, Burban C, Rousselet J, Rossi JP, Branco M, et al. (2011) Incipient allochronic speciation in the pine processionary moth (*Thaumetopoea pityocampa*, Lepidoptera, Notodontidae). *J Evol Biol* 24: 146–158.
31. Friesen VL, Smith AL, Gomez-Diaz E, Bolton M, Furness RW, et al. (2007) Sympatric speciation by allochrony in a seabird. *Proc Natl Acad Sci USA* 104: 18589–18594.
32. Andersson M (1994) Sexual selection. Princeton: Princeton University.
33. Price T, Schluter D (1991) On the low heritability of life-history traits. *Evolution* 45: 853–861.
34. Houle D (1992) Comparing evolvability and variability of quantitative traits. *Genetics* 130: 195–204.
35. West-Eberhard MJ (2003) Developmental plasticity and evolution. Oxford: Oxford University.
36. Thibert-Plante X, Hendry AP (2011) The consequences of phenotypic plasticity for ecological speciation. *J Evol Biol* 24: 326–342.
37. Hagen DW, Gilbertson LG (1972) Geographic variation and environmental selection in *Gasterosteus aculeatus* L. in the Pacific Northwest, America. *Evolution* 26: 32–51.
38. Reimchen TE, Stinson EM, Nelson JS (1985) Multivariate differentiation of parapatric and allopatric populations of threespine stickleback in the Sangan River watershed, Queen Charlotte Islands. *Can J Zool* 63: 2944–2951.
39. Lavin PA, McPhail JD (1993) Parapatric lake and stream sticklebacks on northern Vancouver Island: disjunct distribution or parallel evolution? *Can J Zool* 71: 11–17.
40. Thompson CE, Taylor EB, McPhail JD (1997) Parallel evolution of lake-stream pairs of threespine sticklebacks (*Gasterosteus*) inferred from mitochondrial DNA variation. *Evolution* 51: 1955–1965.
41. Reusch TBH, Wegner KM, Kalbe M (2001) Rapid genetic divergence in postglacial populations of threespine stickleback (*Gasterosteus aculeatus*): the role of habitat type, drainage and geographical proximity. *Mol Ecol* 10: 2435–2445.
42. Hendry AP, Taylor EB (2004) How much of the variation in adaptive divergence can be explained by gene flow? An evaluation using lake-stream stickleback pairs. *Evolution* 58: 2319–2331.
43. Berner D, Adams DC, Grandchamp A-C, Hendry AP (2008) Natural selection drives patterns of lake-stream divergence in stickleback foraging morphology. *J Evol Biol* 21: 1653–1665.
44. Berner D, Roesti M, Hendry AP, Salzburger W (2010) Constraints on speciation suggested by comparing lake-stream stickleback divergence across two continents. *Mol Ecol* 19: 4963–4978.
45. Aguirre WE (2009) Microgeographical diversification of threespine stickleback: body shape-habitat correlations in a small, ecologically diverse Alaskan drainage. *Biol J Linn Soc* 98: 139–151.
46. Deagle BE, Jones FC, Chan YF, Absher DM, Kingsley DM, et al. (2012) Population genomics of parallel phenotypic evolution in stickleback across stream-lake ecological transitions. *Proc R Soc B* 279: 1277–1286.
47. Sharpe DMT, Räsänen K, Berner D, Hendry AP (2008) 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.
48. Berner D, Kaeuffer R, Grandchamp A-C, Raeymaekers JAM, Räsänen K, et al. (2011) Quantitative genetic inheritance of morphological divergence in a lake-stream stickleback ecotype pair: implications for reproductive isolation. *J Evol Biol* 24: 1975–1983.
49. Kaeuffer R, Peichel C, Bolnick DI, Hendry AP (2012) Convergence and non-convergence in ecological, phenotypic, and genetic divergence across replicate population pairs of lake and stream stickleback. *Evolution* 66: 402–418.
50. Roesti M, Hendry AP, Salzburger W, Berner D (2012) Genome divergence during evolutionary diversification as revealed in replicate lake-stream stickleback population pairs. *Mol Ecol* 21: 2852–2862.
51. Hendry AP, Bolnick DI, Berner D, Peichel C (2009) Along the speciation continuum in sticklebacks. *J Fish Biol* 75: 2000–2036.
52. Reimchen TE (1992) Extended longevity in a large-bodied stickleback, *Gasterosteus*, population. *Can Field-Nat* 106: 122–125.
53. Baker JA, Foster SA, Heins DC, Bell MA, King RW (1998) Variation in female life-history traits among Alaskan populations of the threespine stickleback, *Gasterosteus aculeatus* L. (Pisces: Gasterosteidae). *Biol J Linn Soc* 63: 141–159.
54. Baker JA, Cresko WA, Foster SA, Heins DC (2005) Life-history differentiation of benthic and limnetic ecotypes in a polytypic population of threespine stickleback (*Gasterosteus aculeatus*). *Evol Ecol Res* 7: 121–131.
55. Baker JA, Heins DC, King RW, Foster SA (2011) Rapid shifts in multiple life history traits in a population of threespine stickleback. *J Evol Biol* 24: 863–870.
56. Gambling SJ, Reimchen TE (2012) Prolonged life span among endemic *Gasterosteus* populations. *Can J Zool* 90: 284–290.
57. Moodie GEE, Reimchen TE (1976) Phenetic variation and habitat differences in *Gasterosteus* populations of Queen Charlotte Islands. *Syst Zool* 25: 49–61.
58. McPhail JD (1977) Inherited interpopulation differences in size at first reproduction in threespine stickleback, *Gasterosteus aculeatus* L. *Hereditas* 38: 53–60.
59. Nagel L, Schluter D (1998) Body size, natural selection, and speciation in sticklebacks. *Evolution* 52: 209–218.
60. McKinnon JS, Mori S, Blackman BK, David L, Kingsley DM, et al. (2004) Evidence for ecology's role in speciation. *Nature* 429: 294–298.
61. McKinnon JS, Hamel N, Frey N, Chou J, McLeavey L, et al. (2012) Male choice in the stream-anadromous stickleback complex. *PLoS ONE* 7: e37951.
62. Snyder RJ (1991) Quantitative genetic analysis of life histories in two freshwater populations of the threespine stickleback. *Copeia* 2: 526–529.
63. Dufresne F, FitzGerald GJ, Lachance S (1990) Age and size-related differences in reproductive success and reproductive costs in threespine sticklebacks (*Gasterosteus aculeatus*). *Behav Ecol* 1: 140–147.
64. Ishikawa M, Mori S (2000) Mating success and male courtship behaviors in three populations of the threespine stickleback. *Behaviour* 137: 1065–1080.
65. Albert AYK (2005) Mate choice, sexual imprinting, and speciation: a test of a one-allele isolating mechanism in sympatric sticklebacks. *Evolution* 59: 927–931.
66. Boughman JW, Rundle HD, Schluter D (2005) Parallel evolution of sexual isolation in sticklebacks. *Evolution* 59: 361–373.
67. Raeymaekers JAM, Boisjoly M, Delaire L, Berner D, Räsänen K, et al. (2010) Testing for mating isolation between ecotypes: laboratory experiments with lake, stream and hybrid stickleback. *J Evol Biol* 23: 2694–2708.

68. Roesti M, Salzburger W, Berner D (2012) Uninformative polymorphisms bias genome scans for signatures of selection. *BMC Evol Biol* 12: 94.
69. Lucek K, Roy D, Bezaul E, Sivasundar A, Seehausen O (2010) Hybridization between distant lineages increases adaptive variation during a biological invasion: stickleback in Switzerland. *Mol Ecol* 19: 3995–4011.
70. Heller C (1871) Die Fische Tirols und Vorarlbergs. *Zeitschr Ferdinandeum Tir Vor* 16: 295–369.
71. Behrman-Godel J, Gerlach G, Eckmann R (2004) Postglacial colonization shows evidence for sympatric population splitting of Eurasian perch (*Perca fluviatilis* L.) in Lake Constance. *Mol Ecol* 13: 491–497.
72. Jones JW, Hynes HBN (1950) The age and growth of *Gasterosteus aculeatus*, *Pungitius pungitius* and *Spinachia vulgaris*, as shown by their otoliths. *J Anim Ecol* 19: 59–73.
73. Manly BJF (2007) Randomization, bootstrap and Monte Carlo methods in biology. Boca Raton: Chapman & Hall.
74. Rohlf FJ (2001) tpsDig, tpsRelw, tpsRegr. Distributed by the author. Department of Ecology and Evolution, State University of New York, Stony Brook, NY.
75. Hendry AP, Taylor EB, McPhail JD (2002) Adaptive divergence and the balance between selection and gene flow: lake and stream stickleback in the Misty system. *Evolution* 56: 1199–1216.
76. Berner D, Stutz WE, Bolnick DI (2010) Foraging trait (co)variances in stickleback evolve deterministically and do not predict trajectories of adaptive diversification. *Evolution* 64: 2265–2277.
77. Kitano J, Ross JA, Mori S, Kume M, Jones FC, et al. (2009) A role for a neo-sex chromosome in stickleback speciation. *Nature* 461: 1079–1083.
78. Leinonen T, Cano JM, Merilä J (2011) Genetic basis of sexual dimorphism in the threespine stickleback *Gasterosteus aculeatus*. *Heredity* 106: 218–227.
79. Mäkinen HS, Cano JM, Merilä J (2006) Genetic relationships among marine and freshwater populations of the European three-spined stickleback (*Gasterosteus aculeatus*) revealed by microsatellites. *Mol Ecol* 15: 1519–1534.
80. Bentzen P, Ridgway MS, McPhail JD (1984) Ecology and evolution of sympatric sticklebacks (*Gasterosteus*) - spatial segregation and seasonal habitat shifts in the Enos Lake species pair. *Can J Zool* 62: 2436–2439.
81. Bell MA, Foster SA (1994) The evolutionary biology of the threespine stickleback. Oxford: Oxford University.
82. Colosimo PF, Hosemann KE, Balabhadra S, Villareal Jr G, Dickson M, et al. (2005) Widespread parallel evolution in sticklebacks by repeated fixation of ectodysplasin alleles. *Science* 307: 1928–1933.
83. Barrett RDH, Rogers SM, Schluter D (2009) Environment specific pleiotropy facilitates divergence at the ectodysplasin locus in threespine stickleback. *Evolution* 63: 2831–2837.
84. R Development Core Team (2012) R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.
85. Peichel CL, Nereng KS, Oghi KA, Cole BLE, Colosimo PF, et al. (2001) The genetic architecture of divergence between threespine stickleback species. *Nature* 414: 901–905.
86. Coombs JA, Letcher BH, Nislow KH (2008) CREATE: a software to create input files from diploid genotypic data for 52 genetic software programs. *Mol Ecol Res* 8: 578–580.
87. Weir BS, Cockerham CC (1984) Estimating F-statistics for the analysis of population-structure. *Evolution* 38: 1358–1370.
88. Belkhir K, Borsa P, Chikhi L, Raufaste N, Bonhomme F (2004) GENETIX 4.05, logiciel sous Windows pour la génétique des populations. Laboratoire Génome, Populations, Interactions, CNRS UMR 5171, Université de Montpellier II, Montpellier (France). Available: <http://www.genetix.univ-montp2.fr/genetix/genetix.htm>.
89. Hedrick PW (2005) A standardized genetic differentiation measure. *Evolution* 59: 1633–1638.
90. Meirmans PG (2006) Using the AMOVA framework to estimate a standardized genetic differentiation measure. *Evolution* 60: 2399–2402.
91. Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics* 155: 945–959.
92. Hubisz MJ, Falush D, Stephens M, Pritchard JK (2009) Inferring weak population structure with the assistance of sample group information. *Mol Ecol Res* 9: 1322–1332.
93. Earl DA, vonHoldt BM (2011) STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conserv Genet Resour* 4: 359–361.
94. Pritchard J, Wen W (2004) Documentation for structure software: version 2. Available: http://pritch.bsd.uchicago.edu/software/readme_2_1/readme.html.
95. Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol Ecol* 14: 2611–2620.
96. Posada D (2008) jModelTest: Phylogenetic model averaging. *Mol Biol Evol* 25: 1253–1256.
97. Felsenstein J (1981) Evolutionary trees from DNA sequences - a maximum likelihood approach. *J Mol Evol* 17: 368–376.
98. Swofford DL (2003) PAUP*: Phylogenetic analysis using parsimony (* and other methods). Sunderland: Sinauer Associates.
99. Salzburger W, Ewing GB, von Haeseler A (2011) The performance of phylogenetic algorithms in estimating haplotype genealogies with migration. *Mol Ecol* 20: 1952–1963.
100. Kalinowski ST (2011) The computer program STRUCTURE does not reliably identify the main genetic clusters within species: simulations and implications for human population structure. *Heredity* 106: 625–632.
101. Parker GA, Begon M (1986) Optimal egg size and clutch size: effects of environment and maternal phenotype. *Am Nat* 128: 573–592.
102. Bernardo J (1996) The particular maternal effect of propagule size, especially egg size: patterns, models, quality of evidence and interpretations. *Am Zool* 36: 216–236.
103. Thibert-Plante X, Hendry AP (2010) When can ecological speciation be detected with neutral loci? *Mol Ecol* 19: 2301–2314.
104. Bernardo J (1993) Determinants of maturation in animals. *Trends Ecol Evol* 8: 166–173.
105. Berner D, Blanckenhorn WU (2007) An ontogenetic perspective on the relationship between age and size at maturity. *Funct Ecol* 21: 505–512.
106. Craig-Bennett A (1931) The reproductive cycle of the three-spined stickleback, *Gasterosteus aculeatus*, Linn. *Phil Trans R Soc B* 219: 197–279.
107. Baggerman B (1972) Photoperiodic responses in the stickleback and their control by a daily rhythm of photosensitivity. *Gen Comp Endocrinol* 3: 466–476.
108. Herczeg G, Gonda A, Merilä J (2009) Evolution of gigantism in nine-spined sticklebacks. *Evolution* 63: 3190–3200.
109. Shimada Y, Shikano T, Kuparinen A, Gonda A, Leinonen T, et al. (2011) Quantitative genetics of body size and timing of maturation in two nine-spined stickleback (*Pungitius pungitius*) populations. *PLoS ONE* 6.
110. Robinson BW, Wilson DS (1994) Character release and displacement in fishes - a neglected literature. *Am Nat* 144: 596–627.
111. Reimchen TE (1994) Predators and morphological evolution in threespine stickleback. In: Bell MA, Foster SA, editors. *The evolutionary biology of the threespine stickleback*. Oxford: Oxford University. 240–273.
112. Schluter D (1995) Adaptive radiation in sticklebacks: trade-offs in feeding performance and growth. *Ecology* 76: 82–90.
113. Taylor EB, Gerlinsky C, Farrell N, Gow JL (2012) A test of hybrid growth disadvantage in wild, free-ranging species pairs of threespine sticklebacks (*Gasterosteus aculeatus*) and its implications for ecological speciation. *Evolution* 66: 240–251.
114. Gross HP, Anderson JM (1984) Geographic variation in the gillrakers and diet of European sticklebacks, *Gasterosteus aculeatus*. *Copeia* 1: 87–97.
115. Keller O, Krauss E (2000) Die Hydrographie des Bodenseeraums in Vergangenheit und Gegenwart. *Ber St Gallischen Naturwiss Ges* 89: 39–56.
116. Hötzl H (1996) Origin of the Danube-Aach system. *Environ Geol* 27: 87–96.
117. Hagen DW (1967) Isolating mechanisms in threespine sticklebacks (*Gasterosteus*). *J Fish Res Bd Canada* 24: 1637–1692.
118. Gelmond O, von Hippel FA, Christy MS (2009) Rapid ecological speciation in three-spined stickleback *Gasterosteus aculeatus* from Middleton Island, Alaska: the roles of selection and geographic isolation. *J Fish Biol* 75: 2037–2051.
119. Raeymaekers JAM, Van Houdt JKJ, Larmuseau MHD, Geldof S, Volckaert FAM (2007) Divergent selection as revealed by P-ST and QTL-based F-ST in three-spined stickleback (*Gasterosteus aculeatus*) populations along a coastal-inland gradient. *Mol Ecol* 16: 891–905.

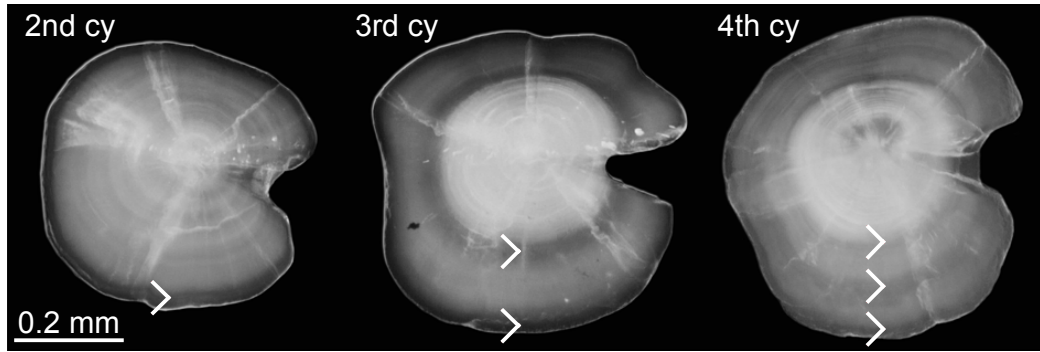
Chapter 9

Chapter 9.1

Repeated lake-stream divergence in stickleback life history within a central European lake basin

Supplementary material

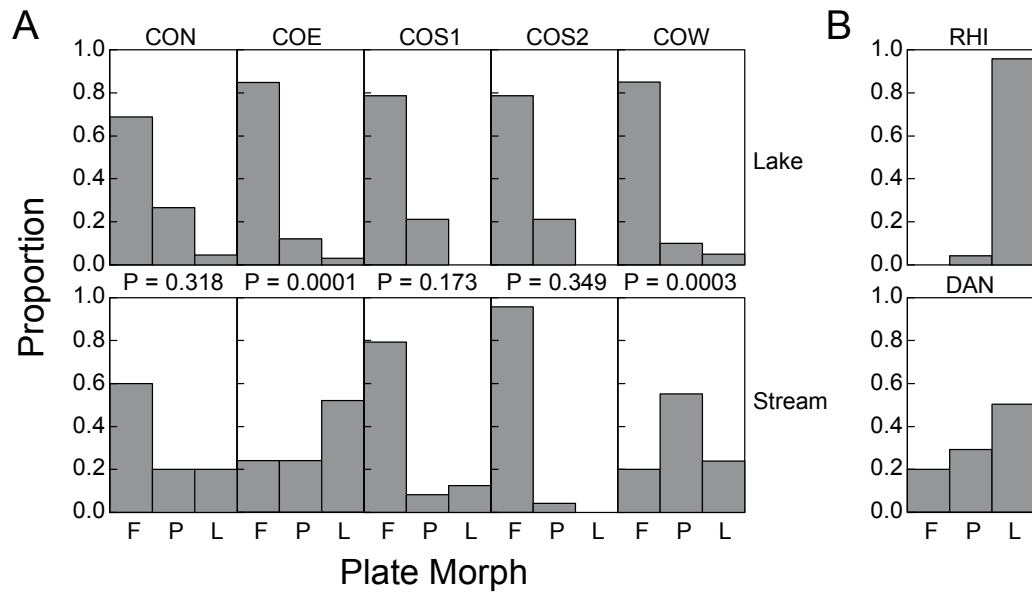
Supporting Tables S1-S3 to Moser et al. (2012) can be obtained online under:
<http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0050620#s5>



Representative sagittal otoliths of stickleback from Lake Constance in their second, third, and fourth calendar year (cy), photographed at 50x magnification. The dark (transparent) ring zones, accreted in spring [72] and used for age determination, are indicated by white arrows.



Representative male stickleback from COW lake (bottom) and COW stream (top). Photo credit: Daniel Berner.



Proportion of fully (F), partially (P), and low-plated (L) stickleback morphs in the five lake-stream pairs (A; lake samples on top, stream samples on the bottom), and in the two solitary stream-resident populations (B). Sample site codes are given in Table 1. P-values are from permutation tests for lake-stream shifts in plate morph frequency within each system. Note the general trend toward plate reduction in the stream samples as compared to the lake samples. P-values and plate morph frequencies for the COW and COS2 system already investigated previously [44] are slightly different from those reported in that study because of random permutation, and because the present study analyzed subsamples of the previous study.

Chapter 9

Chapter 9.2

Depth-dependent abundance of Midas Cichlid fish (*Amphilophus* spp.) in two Nicaraguan crater lakes

Depth-dependent abundance of Midas Cichlid fish (*Amphilophus* spp.) in two Nicaraguan crater lakes

Marie Theres Dittmann · Marius Roesti · Adrian Indermaur · Marco Colombo ·
Martin Gschwind · Isabel Keller · Robin Kovac · Marta Barluenga ·
Moritz Muschick · Walter Salzburger

Received: 4 July 2011 / Revised: 2 February 2012 / Accepted: 5 February 2012 / Published online: 22 February 2012
© Springer Science+Business Media B.V. 2012

Abstract The Midas Cichlid species complex (*Amphilophus* spp.) in Central America serves as a prominent model system to study sympatric speciation and parallel adaptive radiation, since small arrays of equivalent ecotype morphs have evolved independently

in different crater lakes. While the taxonomy and evolutionary history of the different species are well resolved, little is known about basic ecological parameters of Midas Cichlid assemblages. Here, we use a line transect survey to investigate the depth-dependent abundance of *Amphilophus* spp. along the shores of two Nicaraguan crater lakes, Apoyo and Xiloá. We find a considerable higher density of Midas cichlids in Lake Xiloá as compared to Lake Apoyo, especially at the shallowest depth level. This might be due to the higher eutrophication level of Lake Xiloá and associated differences in food availability, and/or the presence of a greater diversity of niches in that lake. In any case, convergent forms evolved despite noticeable differences in size, age, eutrophication level, and carrying capacity. Further, our data provide abundance and density estimates for Midas Cichlid fish, which serve as baseline for future surveys of these ecosystems and are also relevant to past and future modeling of ecological speciation.

Marie Theres Dittmann and Marius Roesti contributed equally to this study.

Handling editor: Christian Sturmbauer

Electronic supplementary material The online version of this article (doi:10.1007/s10750-012-1024-1) contains supplementary material, which is available to authorized users.

M. T. Dittmann · M. Roesti · A. Indermaur ·
M. Colombo · M. Gschwind · I. Keller ·
R. Kovac · M. Muschick · W. Salzburger (✉)
Zoological Institute, University of Basel, Vesalgasse 1,
4051 Basel, Switzerland
e-mail: walter.salzburger@unibas.ch

Present Address:

M. T. Dittmann
Group of Animal Nutrition, Institute of Agricultural
Science, Swiss Federal Institute of Technology,
Universitätsstrasse 2, 8092 Zurich, Switzerland

I. Keller
IFM-Geomar, Evolutionsbiologie, Düstenbrooker Weg
20, 24105 Kiel, Germany

M. Barluenga
Museo Nacional de Ciencias Naturales, CSIC, José
Gutierrez Abascal 2, 28006 Madrid, Spain

Keywords Sympatric speciation ·
Parallel adaptive radiation · Fish density estimates ·
Crater Lake Apoyo · Crater Lake Xiloá · Ecology

Introduction

The species flocks of cichlid fishes in the East African Great Lakes Victoria, Malawi, and Tanganyika are prime model systems in evolutionary biology and, particularly, in research focusing on speciation,

adaptive radiation, and parallel evolution (reviewed in Kocher, 2004; Salzburger, 2009; Sturmbauer et al., 2011). One of the most outstanding features of the East African cichlid assemblages is their species richness, with each of the Great Lakes harboring hundreds of endemic species. The downside of this unparalleled diversity is that these species flocks are notoriously difficult to study in their entirety, which makes it attractive to study simpler cichlid communities in smaller water bodies. In the last years surveys of crater lakes cichlids proved especially fruitful, mostly due to the degree of isolation of their cichlid assemblages (Schliewen et al., 1994; Barluenga & Meyer, 2004; Barluenga et al., 2006). The probably best-studied cichlids in volcanic crater lakes belong to the Midas Cichlid species complex (*Amphilophus* spp.), which is native to Central America. Midas cichlids are abundant in the large lakes of Nicaragua (Lake Nicaragua and Lake Managua) and associated rivers in Nicaragua and northern Costa Rica. Interestingly, Midas Cichlids have also colonized various volcanic crater lakes in the area (Barlow, 1976; Barluenga & Meyer, 2004, 2010), which emerge when calderas of extinct volcanoes of the ‘Pacific Ring of Fire’ become filled with water.

This study focuses on the *Amphilophus* assemblages in two of these crater lakes, Apoyo and Xiloá, which contain two independent, yet ecologically and morphologically very similar sets of Midas cichlid species (Elmer et al., 2010; Geiger et al., 2010a). The lakes are similar in some aspects, such as their volcanic origin, but they do differ in others (Barlow, 1976; Sussman, 1985; Waid et al., 1999; McKaye et al., 2002; Barluenga & Meyer, 2010): With a surface area of 21.1 km² and a maximum depth of 142 m, Lake Apoyo is larger and deeper than Lake Xiloá, which has a surface area of 3.8 km² and a maximum depth of 89 m (Table 1). Also, compared to the nutrient-rich Lake Xiloá, Lake Apoyo is oligotrophic. Furthermore, they differ in the number of cichlid species. Crater Lake Apoyo is suggested to harbor six endemic species of the *Amphilophus* complex (Barlow, 1976; Stauffer et al., 2008; Geiger et al., 2010b) (Supplementary Table 1), which most likely go back to a seeding lineage from adjacent Lake Nicaragua (Barluenga et al., 2006); together with *Parachromis managuense* and the recently introduced African species *Oreochromis aureus* and *O. niloticus*, these are the only cichlids found in this lake. In Lake Xiloá three to four endemic species of the *Amphilophus* species complex are described (McKaye et al., 2002;

Table 1 General descriptors of size, depth, age, visibility, fish density, and population size of the crater lakes Apoyo and Xiloá

	Apoyo	Xiloá
Surface area (km ²)	21.1 ^a	3.8 ^a
Maximum depth (m)	142 ^a	89 ^a
Age (year)	<23.000 ^a	ca. 10.000 ^a
Secchi depth (m)	5–7	3
Cichlid density along shore (individuals per 10 m transect)	11.3	19.9
Total number of <i>Amphilophus</i> spp. along shore (estimated)	83.000	66.000

^a Barluenga & Meyer (2010)

Stauffer & McKaye, 2002) (Supplementary Table 1), which derive from the close-by Lake Managua stocks (Barluenga & Meyer, 2010). In addition to the Midas Cichlid fish, Lake Xiloá is inhabited by eight additional cichlid species, which either migrated naturally from nearby Lake Managua, or were introduced by humans, as might be the case for *Parachromis managuense* (Kullander & Hartel, 1997).

Here, we present a comparative study of cichlid abundance and density estimates in the two Central American calderas Lake Apoyo and Lake Xiloá. The set-up consisting of two rather similar crater lakes seeded independently by more or less the same ancestral line that subsequently radiated in parallel appears ideal to disentangle the biotic and abiotic factors influencing parallel adaptive radiation, particularly in its early stages. Many adaptive radiations appear to proceed in discrete stages starting with an initial diversification into macrohabitats (Streelman & Danley, 2003; Gavrillets & Losos, 2009), which—in fishes—is often associated with differentiation along the benthic-limnetic (pelagic) axis (Schluter & McPhail, 1992; Gíslason et al., 1999; Barluenga et al., 2006; Rutschmann et al., 2011). That independent adaptive radiations of the same group of organisms in similar ecological settings often result in similar morphologies is generally taken as strong evidence for natural selection (and the importance of ecology in speciation) (see Schluter & Nagel, 1995; Losos et al., 1998). On the other hand, the degree of similarity observed in convergent species pairs of cichlids has led some authors to question whether natural selection alone is sufficient to produce such matching morphologies, or whether genetic or developmental constraints have

contributed to the evolution of convergent forms (see, e.g., Brakefield, 2006). Even in the genomic era it is difficult to determine the relative contribution of natural selection and developmental channeling to parallel evolution. One possibility is to apply genetic and genomic experiments (reviewed in: Brakefield, 2006; Arendt & Reznick, 2008). In addition, one should inspect parallel radiations with respect to key ecological parameters. Under the assumption that ecology is the driving force behind parallel adaptive radiation, it is expected that not only the outcome of the radiations should be the same, but that the radiations should also follow the same steps and should show the same (ecological) characteristics. In the case of the parallel radiations of the Midas Cichlid in crater lakes Apoyo and Xiloá

and Xiloá, the outcome in form of morphologically equivalent species is obviously quite similar (Fig. 1) and there is evidence that the radiations progressed in a similar fashion (Barluenga et al., 2006; Barluenga & Meyer, 2010; Elmer et al., 2010). It is not known, however, whether the communities in the seemingly similar crater lakes Apoyo and Xiloá are also similar in terms of ecological parameters such as fish densities and depth distributions.

In this study, we applied transect surveys to record the abundance of *Amphilophus* spp. in crater lakes Apoyo and Xiloá. Applying SCUBA diving and snorkeling, fish were counted at different locations and depth levels to provide data on densities of cichlids in both lakes. We hypothesized that the

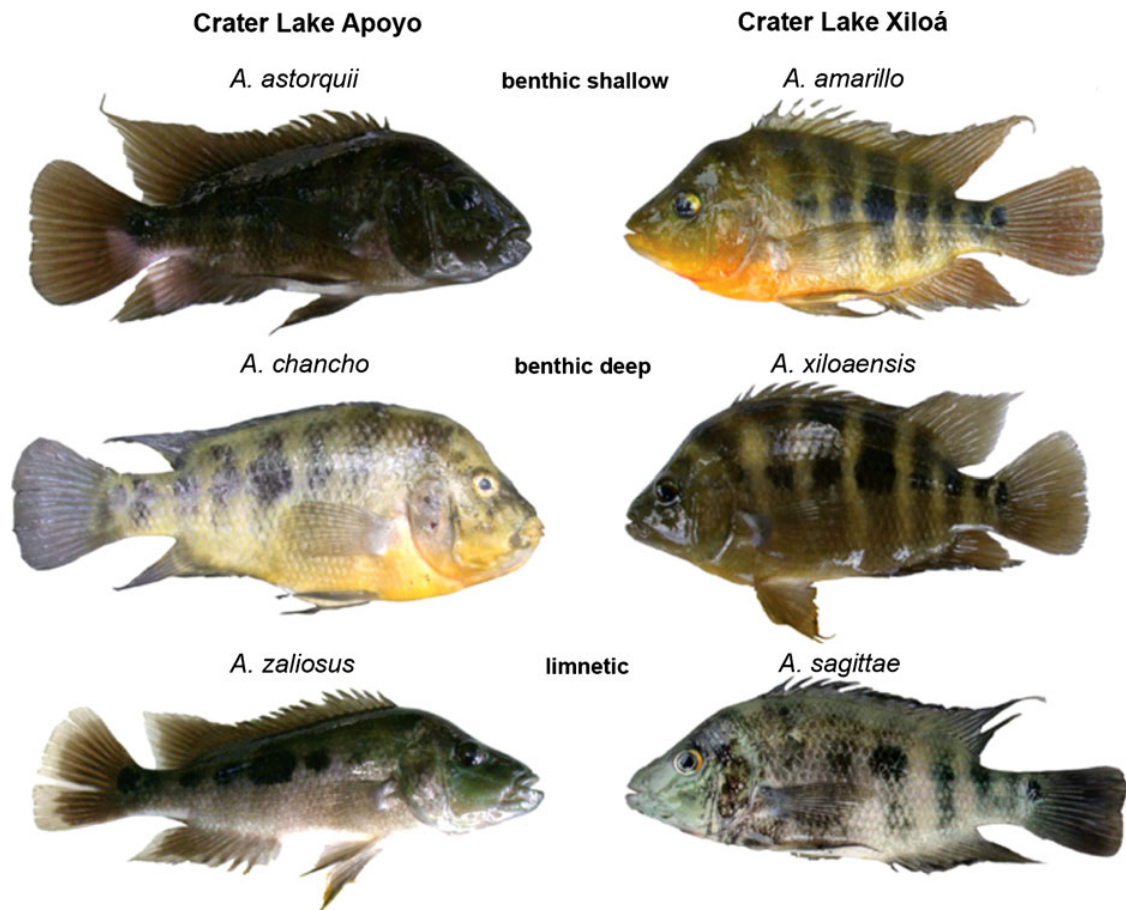


Fig. 1 Convergent phenotypes that evolved independently in the two Nicaraguan crater lakes Apoyo and Xiloá. Three species pairs are shown: benthic species using the shallow areas of the

lakes; benthic species using the deeper areas of the lakes; and limnetic species inhabiting the open water column

density and distribution of Midas cichlids should be rather similar in both crater lakes due to their similar mode of origin and structure. In addition, this study aims to add ecological data in the form of abundance estimates for *Amphilophus* spp. to theoretical studies on sympatric and/or ecological speciation. Gavrillets et al. (2007), for example, investigated under which biological conditions rapid colonization of a new niche followed by sympatric or parapatric speciation in Lake Apoyo is theoretically possible. However, in their models, Gavrillets et al. (2007) were lacking empirical data on several important biological parameters (including abundance estimates). Finally, knowledge of the natural abundance of a population, species, or species group is fundamental not only to biological research but also to the management of wildlife populations. This is important in the case of crater lakes Apoyo and Xiloá, too, where cichlid fishes make up the main fraction of the ichthyofauna and provide a valuable food resource for local people (Schuster, 1957; Lin, 1961; Barlow, 1976). Importantly, through the recent introduction of African tilapiine cichlid species (*Oreochromis* spp.), the endemic cichlids of Lake Apoyo are thought to be threatened (McKaye et al., 1995; McCrary et al., 2001; Barluenga & Meyer, 2004), calling for an evaluation of the conservation status of the endemic faunas in the two crater lakes. Our data should, thus, provide important baseline references, with which upcoming impacts on the native cichlid abundance can be assessed.

Materials and methods

Study area and period

Field work was carried out in the two crater lakes Apoyo and Xiloá in Nicaragua, Central America, in September 2009. Diving was performed during the day by almost invariably good weather conditions. At the time of the study, water temperatures ranged between 29 and 31°C on all surveyed depth levels in both lakes. Transect sites were chosen randomly in both lakes, balanced, however, for different geographical locations within each lake (Supplementary Table 2). As crater lakes have a relatively homogeneous habitat structure, the transects are representative of the habitat composition in each lake.

Transect surveys

We used fish counts along line transects to compare the depth-dependent abundance and density of *Amphilophus* spp. between the two lakes. Six transects were studied in the larger Lake Apoyo and four transects in the smaller Lake Xiloá. The start and end coordinates of each transect were taken with a handheld GPS from a boat (Supplementary Table 2). Depth levels at 10, 15, and 20 m were covered for each transect by a SCUBA diving buddy pair, whereas the 5 m depth level was covered by snorkelers (whenever the visibility was sufficient).

Transect length was determined by the distance covered during 10–15 min of diving (depending on the available air). Diving pace was moderate but varied between transects according to visibility and the quantity of fish that had to be counted, leading to variation in the lengths of the different transects. After having covered a transect one way, buddy pairs remained at their set depth level for 10 min to leave enough time for the fish to restore an undisturbed distribution. The end of each transect was marked with a buoy, which enabled the recording of the GPS coordinates. Buddy pairs then returned along the line transect back to the starting point. Diving was performed at 2 m above the substrate whereby dive buddies were swimming beside each other, individually counting all *Amphilophus* spp. individuals larger than ca. 5 cm within a visual field of about 4 m distance and 2 m to either side of the transect line. Snorkelers covering the 5 m depth used the same method and tried to remain at a depth of 3 m as much as possible. Owing to the difficulty to clearly identify species in sub-adult or non-breeding life stages underwater and the ongoing debate and steady changes in species classification, the overall number of *Amphilophus* spp. individuals was counted and no attempts were made to distinguish species, hybrids, or morphotypes (e.g., Barlow, 1976; McKaye et al., 2002; Bunje et al., 2007; Stauffer et al., 2008). In this visual survey a minimal bias among and within observers is expected due to individual survey differences (Thompson & Mapstone, 1997). To remove such potential confounding effects, observers alternated between different depth levels and in buddy pair partners at consecutive transects. The total number of dives over all transects was 36 (including each two persons diving back and forth), resulting in 144 single transect records.

In addition, Secchi depth measurements were taken from a boat to determine the water transparency at several random locations in both lakes.

Data analysis

To determine the average number of *Amphilophus* spp. individuals for every transect at each depth level separately, we averaged the fish counts by the two buddy team partners including the replicates from diving back and forth (Supplementary Table 3). We then calculated the average numbers of individuals per 10 m transect length for each depth level for every transect (Fig. 2), which we tested for normal distribution by applying a Shapiro–Wilk test. Using this data we tested for an overall difference in the density of *Amphilophus* spp. between lakes using Mann–Whitney U tests. We further applied a linear mixed model (LMM, LME4 package, Bates et al., 2011) to test for a difference in number and depth-distribution of individuals between the lakes by including the number of individuals counted per 10 m as the dependent variable, and lake and depth level as predictors. Assumptions of the LMM were visually checked. Since we assumed a potential difference in the depth-distribution of individuals between lakes, we included the interaction of lake and depth in the model. Furthermore, to correct for dependence in our data, we included transect as random factor. To further explore the data for effects not captured by the LMM, we applied separate Mann–Whitney U Tests for each depth level to test for depth-dependent differences in fish abundance between lakes. To roughly estimate the

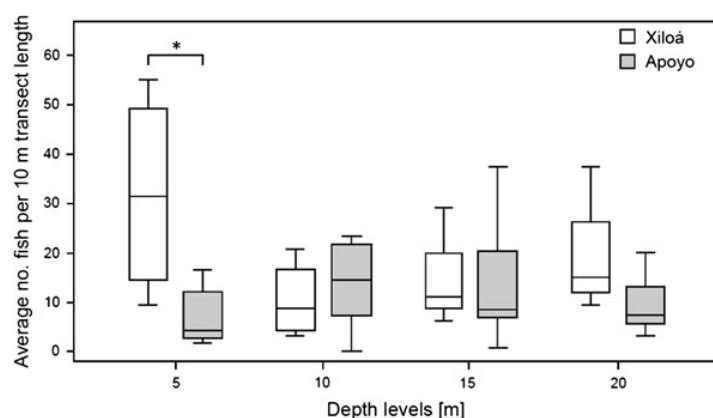
total number of Midas cichlids for both lakes, the numbers of fish per 10 m were extrapolated to the total circumference of the lake. This was calculated by summing up the average number of individuals at all four depth levels (Suppl. Table 3) multiplied by the circumference of the lake. All analyses was performed using R 2.9.2 (R Foundation for Statistical Computing, Vienna, Austria).

Results

The average number of *Amphilophus* spp. individuals per 10 m transect length in Apoyo across all transects and depth levels was 11.3 (min = 0, max = 37, SD = 9.5), which did not differ significantly from Lake Xiloá with 19.9 fish per 10 m transect length (min = 3, max = 55, SD = 15.7) (Mann–Whitney U test, $N = 36$, $p = 0.112$). The LMM did not reveal a significant interaction between lake and depth ($t = 0.1692$, $p = 0.169$) (Fig. 2). However, testing for single depth levels between the lakes revealed a marginally significant difference at the 5 m depth level (Mann–Whitney U test, $N = 10$, $W = 18$, $p = 0.050$). The pairwise comparison of numbers of fish per 10 m transect at the other depth levels exhibited no significant difference between the lakes (Mann–Whitney U test, 10 m: $N = 10$, $p = 0.394$; 15 m: $N = 10$, $p = 0.796$; 20 m: $N = 8$, $p = 0.180$).

Extrapolating the average number of *Amphilophus* spp. individuals of all transects and depth levels to the total circumference in both lakes (Apoyo approx. 18.2 km; Xiloá approx. 8.3 km) revealed a similar

Fig. 2 Average number of *Amphilophus* spp. individuals per 10 m transect at each depth level for Lake Xiloá and Lake Apoyo. “*” denotes a marginally significant difference in cichlid fish density between the lakes (Mann–Whitney U test, $N = 10$, $p = 0.050$)



total number of fish in both lakes along the shoreline: ca. 83,000 individuals (13,000 to 150,000) in Lake Apoyo and ca. 66,000 individuals (13,000 to 120,000) in Lake Xiloá.

The Secchi depth, measured randomly several times in both lakes, ranged between 5 and 7 m in Lake Apoyo, compared to an approximately constant Secchi depth of 3 m in Lake Xiloá.

Discussion

Benefits of fish abundance estimates are diverse. The comparison of fish abundances between comparable ecosystems (e.g., between lakes) that differ in only few and well-defined ecological factors, allows to draw general conclusions on the possible impact of these factors on fish abundances and the composition and evolution of communities. This is especially the case when members of the same lineage radiated in parallel. Furthermore, in conservation biology and wildlife management, for example, changes in abundance of a fish species or population in a specific area may give an estimate for its “ecological health”. This allows to define appropriate conservation strategies as well as to evaluate the (long-term) effects of habitat or species-specific conservation actions (Cheal & Thompson, 1997; Witmer, 2005). To estimate the impact of naturally induced (e.g., by a hurricane) or human-induced (e.g., by industrial fishery) changes on fish abundance, a baseline abundance needs to be established against which future levels of impact can be assessed (Jennings & Blanchard, 2004; Silvano et al., 2009). Then, abundance estimates are valuable to evaluate the relative importance and status of a fish species in an ecosystem, such as in a predator–prey relationship in the food web. Finally, mathematical modeling in fields such as evolutionary biology provides more accurate, theoretical insights into biological processes. Most often, however, theoretical approaches lack data from empirical work such as abundance estimates that would allow to make biologically reasonable assumptions and to apply mathematical models to particular case studies (see, e.g., Gavrilets et al., 2007).

The above reasons have been the motivation for this comparative study of Midas cichlid fish (*Amphilophus* spp.) abundance and density estimates in the two comparable Nicaraguan crater lakes, Apoyo and

Xiloá. Despite the lack of statistical significance, our data reveal an almost twofold higher density of cichlid fish along the shoreline in Lake Xiloá as compared to Lake Apoyo. At a depth of 5 m, we found a more than fourfold higher density of Midas cichlids in Lake Xiloá (Fig. 2). Overall, however, as a consequence of the higher density of fish in the smaller lake Xiloá, the absolute numbers of *Amphilophus* spp. are relatively similar in both lakes—at least along the shore habitat covered by our survey.

Differences in food availability could explain the different densities of *Amphilophus* spp. between the two crater lakes. Indeed, the two lakes differ in their level of eutrophication: Lake Apoyo is an oligotrophic environment, whereas Lake Xiloá is relatively more eutrophic. But why would higher fish densities then only be found at shallow areas and not throughout Lake Xiloá? Eutrophication leads to a considerable reduction of ambient light at deeper waters (e.g. Koch, 2001), which can restrict photosynthesis to the shallow waters where sufficient ambient light is available for primary production (see Secchi depth in Table 1). This can directly (e.g., algae-feeders) or indirectly (e.g., through the food web) lead to higher fish densities in the shallow area. Higher fish densities in more turbid waters may also be explained by the reduced performance of predators, such as birds, which under turbid conditions have more difficulties to spot fish. It has previously been shown that reduced visibility can influence color-recognition in cichlids, and, hence, may have an impact on intraspecific (and interspecific) species recognition and communication (see, e.g., Seehausen, 1997, 2008). Whether this is also the case in Nicaraguan crater lakes remains to be tested.

An alternative explanation for the higher density of cichlids in Lake Xiloá could be the availability of ecologically more diverse niches in this lake, e.g., in the shallow area where differences in the densities of *Amphilophus* spp. are greatest. This could also explain the higher variance in fish counts at the 5 m depth level in Lake Xiloá compared to the other depth levels. Perhaps it is a combination of both factors, eutrophication and habitat complexity, that leads to higher fish densities in Lake Xiloá. A more thorough analysis of the habitat structure would be necessary to clarify this point. Furthermore, there is no knowledge on fish densities in deeper and open waters, which would allow a comprehensive comparison of both lakes. Such fish counts at deeper waters seem particularly

interesting, since we observed a distinct and clear water layer below a depth of 35 m in Lake Xiloá.

Crater lakes Apoyo and Xiloá are inhabited by a similar set of convergent *Amphilophus* ecotype morphs (Fig. 1) making the Midas Cichlid complex an ideal system to study parallel evolution (see, e.g., McKaye et al., 2002; Barluenga et al., 2006; Elmer et al., 2010). While taxonomy, morphology, and evolutionary history of the species complex is largely resolved (see Barluenga et al., 2006; Barluenga & Meyer, 2010; Elmer et al., 2010; Geiger et al., 2010a, b), little is known about basic ecological parameters such as the relative densities of the different species. Our study is the first to provide such data. We uncover a rather similar overall number of *Amphilophus* spp. individuals in both lakes, but also account differences in densities, especially in the shallow area (see above). Interestingly, the shallow areas of Lake Xiloá are not only characterized by larger densities of Midas cichlids, but also by the presence of additional cichlid species (see Supplementary Table 1). It remains unclear whether these never arrived in Lake Apoyo (e.g., because of the larger distance to a large lake), or whether these could not establish themselves there (e.g., because of the eutrophic situation). In any case, convergent phenotypes evolved in both crater lakes despite noticeable differences in size and age of the respective lake (see Table 1), in community structure (the presence/absence of other cichlid species; Supplementary Table 1), and in fish densities (Fig. 2). This corroborates the view that the initial steps of ecological speciation in fish species flocks follow similar pathways in form of a splitting into benthic and limnetic types (see, e.g., Schluter & McPhail, 1992; Salzburger, 2009), which does not seem to be dependent on phylogenetic background and parameters such as size or age of a lake or level of eutrophication. Apparently, it is enough that a benthic-limnetic axis is present in a lake (see Barluenga et al., 2006).

The Midas cichlid fauna from Lake Apoyo represents one of the most famous examples for sympatric speciation (Barluenga et al., 2006), and has attracted theoretical modeling work. Gavrillets et al. (2007), for example, investigated whether at all and under which ecological conditions sympatric speciation is likely to have occurred in lake Apoyo. One of the parameters incorporated into the model of Gavrillets et al. (2007) was the carrying capacity (K) of Lake Apoyo. Carrying capacity stands for the maximum number

of individuals that can live in a particular environment given the available nutrients and without causing detrimental effects. Gavrillets et al. (2007) concluded that intermediate carrying capacities ($K = 16.000$) are propensive for sympatric speciation, whereas large carrying capacities ($K = 32.000$ – 51.200 , depending on the model) would rather lead to the evolution of a single, generalistic species. Our estimates of K (ca. 83.000 and ca. 66.000 individuals in Lakes Apoyo and Xiloá, respectively) lie above these numbers, although these estimates refer to counts at four depth levels along the shoreline only and nothing is known about fish densities below 20 m. One also has to consider that Gavrillets et al. (2007) assumed the presence of a single age class (i.e., generation) at a given time. Our counts certainly included members from different age classes, although we lack detailed information on age distribution. Taken together, the carrying capacities assumed by Gavrillets et al. (2007) to model sympatric speciation in Lake Apoyo seem to be slightly—however not substantially—underestimated compared to our findings and it would now be interesting to evaluate what effect this has on available models.

Although a reproducing population of invasive *Oreochromis* spp. (tilapias) has been reported for Lake Apoyo in previous studies (McKaye et al., 1995; McCrary et al., 2001), we did not observe any tilapiine species during our fieldwork. These African cichlids were reported to feed on stonewort beds (*Chara* spp.) and are likely to account for the temporal elimination of these algae in Lake Apoyo (McKaye et al., 1995; McCrary et al., 2001, Canonico et al., 2005). However, we found extensive stonewort beds in Lake Apoyo. This suggests that tilapia populations might have failed to establish permanently in an oligotrophic environment such as Lake Apoyo.

Conclusions

Our study gives estimates of cichlid fish densities in two crater lakes in Nicaragua, Apoyo and Xiloá. We find that parallel ecotype morphs evolved despite noticeable differences in size, age, eutrophication level, and carrying capacity. We provide ecological data for understanding the carrying capacity of the systems in order to apply it to modeling sympatric/parapatric speciation. Furthermore, it sets baseline abundance estimates for cichlid fish in Nicaraguan

crater lakes, to which future ecological health assessments of these lakes can be compared.

Acknowledgments We are grateful to C. Heule, N. Hue and A. Theis for assisting us with diving; B. Christ and T. Suter for their helping hand at dive sites; the Ministerio del Ambiente y los Recursos Naturales Nicaragua (MARENA) for research permits; E. P. van den Berghe for logistical help and scientific expertise; and two anonymous referees, Associated Editor C. Sturmbauer and the Editor K. Martens for valuable comments on the manuscript; T. Roth for statistical support, and the “Fuerzas Armadas de Nicaragua” for boat cruises and air supply. This project was funded by grants from the European Research Council (ERC) and the Swiss NSF.

References

- Arendt, J. & D. Reznick, 2008. Convergence and parallelism reconsidered: what have we learned about the genetics of adaptation? *Trends in Ecology & Evolution* 23: 26–32.
- Barlow, G. W., 1976. The Midas cichlid in Nicaragua. In Thorson, T. B. (ed.), *Investigations of the Ichthyofauna of Nicaraguan Lakes*, University of Nebraska-Lincoln: 333–358.
- Barluenga, M. & A. Meyer, 2004. The Midas cichlid species complex: incipient sympatric speciation in Nicaraguan cichlid fishes. *Molecular Ecology* 13: 2061–2076.
- Barluenga, M. & A. Meyer, 2010. Phylogeography, colonization and population history of the Midas cichlid species complex (*Amphilophus* spp.) in the Nicaraguan crater lakes. *BMC Evolutionary Biology* 10: 326.
- Barluenga, M., K. N. Stölting, W. Salzburger, M. Muschick & A. Meyer, 2006. Sympatric speciation in Nicaraguan Crater Lake cichlid fish. *Nature* 439: 719–723.
- Bates, D., M. Mächler & B. Bolker, 2011. LME4: linear mixed-effects model using S4 classes. R package, v. 0.999375-40. <http://cran.r-project.org>.
- Brakefield, P. M., 2006. Evo-devo and constraints on selection. *Trends in Ecology & Evolution* 21: 362–368.
- Bunje, P. M., M. Barluenga & A. Meyer, 2007. Sampling genetic diversity in the sympatrically and allopatrically speciating Midas cichlid species complex over a 16 year time series. *Bmc Evolutionary Biology* 7: 25.
- Canonico, G. C., A. Arthington, J. K. McCrary & M. L. Thieme, 2005. The effects of introduced tilapias on native biodiversity. *Aquatic Conservation: Marine and Freshwater Ecosystems* 15: 463–483.
- Cheal, A. J. & A. A. Thompson, 1997. Comparing visual counts of coral reef fish: implications of transect width and species selection. *Marine Ecology Progress Series* 158: 241–248.
- Elmer, K. R., H. Kusche, T. K. Lehtonen & A. Meyer, 2010. Local variation and parallel evolution: morphological and genetic diversity across a species complex of Neotropical Crater Lake cichlid fishes. *Philosophical Transactions of the Royal Society London B* 365: 1763–1782.
- Gavrilets, S., A. Vose, M. Barluenga, W. Salzburger & A. Meyer, 2007. Case studies and mathematical models of ecological speciation. 1. Cichlids in a crater lake. *Molecular Ecology* 16: 2893–2909.
- Gavrilets, S. & J. B. Losos, 2009. Adaptive radiation: contrasting theory with data. *Science* 323: 732–737.
- Geiger, M. F., J. K. McCrary & U. K. Schliewen, 2010a. Not a simple case – A first comprehensive phylogenetic hypothesis for the Midas cichlid complex in Nicaragua (Teleostei: Cichlidae: Amphilophus). *Molecular Phylogenetics and Evolution* 56: 1011–1024.
- Geiger, M. F., J. K. McCrary & J. R. Stauffer, 2010b. Description of two new species of the Midas cichlid complex (Teleostei: Cichlidae) from Lake Apoyo, Nicaragua. *Proceedings of the Biological Society of Washington* 123: 159–173.
- Gíslason, D., M. M. Ferguson, S. Skúlason & S. S. Snorrason, 1999. Rapid and coupled phenotypic and genetic divergence in Icelandic Arctic charr (*Salvelinus alpinus*). *Canadian Journal of Fisheries and Aquatic Sciences* 56: 2229–2234.
- Jennings, S. & J. Blanchard, 2004. Fish abundance with no fishing: predictions based on macroecological theory. *Journal of Animal Ecology* 73: 632–642.
- Koch, E. W., 2001. Beyond light: physical, geological and geochemical parameters as possible submersed aquatic vegetation habitat requirements. *Estuaries* 24: 1–17.
- Kocher, T. D., 2004. Adaptive evolution and explosive speciation: the cichlid fish model. *Nature Reviews Genetics* 5: 288–298.
- Kullander, S. & K. E. Hartel, 1997. The systematic status of cichlid genera described by Louis Agassiz in 1859: *Amphilophus*, *Baiodon*, *Hypsophrys* and *Parachromis* (Teleostei: Cichlidae). *Ichthyological Explorations of Freshwaters* 7: 193–202.
- Lin, S. Y., 1961. Informe al goberino de Nicaragua sobre el desarrollo de un proyecto de pesquerías continentals en dicho país. Informe F. A. O. 1347: 1–21.
- Losos, J. B., T. R. Jackman, A. Larson, K. Queiroz & L. Rodriguez-Schettino, 1998. Contingency and determinism in replicated adaptive radiations of island lizards. *Science* 279: 2115–2118.
- McCrary, J. K., E. P. van den Berghe, K. R. McKaye & L. J. Lopez Perez, 2001. Tilapia cultivation: a threat to native species in Nicaragua. *Encuentro* 58: 9–19.
- McKaye, K. R., J. D. Ryan, J. R. Stauffer, L. J. Lopez Perez & E. P. van den Berghe, 1995. African tilapia in Lake Nicaragua: ecosystem in transition. *BioScience* 45: 406–411.
- McKaye, K. R., J. R. Stauffer, E. P. van den Berghe, R. Vivas, L. J. Lopez Perez, J. K. McCrary, R. Waid, A. Konings, W. J. Lee & T. D. Kocher, 2002. Behavioral, morphological and genetic evidence of divergence of the Midas Cichlid species complex in two Nicaraguan crater lakes. *Cuadernos de Investigación de la UCA* 12: 19–47.
- Rutschmann, S., M. Matschiner, M. Damerau, M. Muschick, M. F. Lehmann, R. Hanel & W. Salzburger, 2011. Parallel ecological diversification in Antarctic notothenioid fishes as evidence for adaptive radiation. *Molecular Ecology* 20: 4707–4721.
- Salzburger, W., 2009. The interaction of sexually and naturally selected traits in the adaptive radiations of cichlid fishes. *Molecular Ecology* 18: 169–185.
- Schliewen, U. K., D. Tautz & S. Paabo, 1994. Sympatric speciation suggested by monophyly of crater lake cichlids. *Nature* 368: 629–632.
- Schluter, D. & J. D. McPhail, 1992. Ecological character displacement and speciation in sticklebacks. *American Naturalist* 140: 85–108.

- Schluter, D. & L. Nagel, 1995. Parallel speciation by natural selection. *American Naturalist* 140: 292–301.
- Schuster, W. H., 1957. Informe al goberino de Nicaragua sobre fomento de la pesca continental. Informe F. A. O. 607: 1–13.
- Seehausen, O., 1997. Cichlid fish diversity threatened by eutrophication that curbs sexual selection. *Science* 277: 1808–1811.
- Seehausen, O., Y. Terai, I. S. Magalhaes, K. L. Carleton, H. D. J. Mrosso, R. Miyagi, I. van der Sluijs, et al., 2008. Speciation through sensory drive in cichlid fish. *Nature* 455: 620–626.
- Silvano, R. A. M., M. Ramires & J. Zuanon, 2009. Effects of fisheries management on fish communities in the floodplain lakes of a Brazilian Amazonian Reserve. *Ecology of Freshwater Fish* 18: 156–166.
- Stauffer, J. R. & K. R. McKaye, 2002. Descriptions of three new species of cichlid fishes (Teleostei: Cichlidae) from Lake Xiloá, Nicaragua. *Cuadernos de Investigación de la Universidad Centroamericana* 12: 1–18.
- Stauffer, J. R., J. K. McCrary & K. E. Black, 2008. Three new species of cichlid fishes (Teleostei: Cichlidae) from Lake Apoyo, Nicaragua. *Proceedings of the Biological Society Washington* 121: 117–129.
- Streelman, J. T. & P. D. Danley, 2003. The stages of vertebrate evolutionary radiation. *Trends in Ecology & Evolution* 18: 126–131.
- Sturmbauer, C., M. Husemann & P. D. Danley, 2011. Explosive speciation and adaptive radiation of East African cichlid fishes. In Zachos, F. E. & J. C. Habel (eds.), *Biodiversity Hotspots: Distribution and Protection of Conservation Priority Areas*. Springer: 333–362.
- Sussman, D., 1985. Apoyo caldera, Nicaragua: a major quaternary silicic eruptive center. *Journal of Volcanology and Geothermal Research* 24: 249–282.
- Thompson, A. A. & B. D. Mapstone, 1997. Observer effects and training in underwater visual surveys of reef fishes. *Marine Ecology Progress Series* 154: 53–63.
- Waid, R. M., R. L. Raesly, K. R. McKaye & J. K. McCrary, 1999. Zoogeografía íctica de lagunas cratéricas de Nicaragua. *Encuentro* 51: 65–80.
- Witmer, G. W., 2005. Wildlife population monitoring: some practical considerations. *Wildlife Research* 32: 259.

Chapter 9

Chapter 9.2

Depth-dependent abundance of Midas Cichlid fish (*Amphilophus* spp.) in two Nicaraguan crater lakes

Supplementary material

Supplementary Material

Supplementary Table 1: Cichlid fish diversity in lakes Apoyo and Xiloá.

Lake Apoyo – Midas cichlid species (endemic)
<i>Amphilophus zaliosus</i> Barlow and Munsey 1976
<i>Amphilophus flaveolus</i> Stauffer <i>et al.</i> 2008
<i>Amphilophus chancho</i> Stauffer <i>et al.</i> 2008
<i>Amphilophus astorquii</i> Stauffer <i>et al.</i> 2008
<i>Amphilophus globosus</i> Geiger <i>et al.</i> 2010
<i>Amphilophus supercilius</i> Geiger <i>et al.</i> 2010
Lake Apoyo – other cichlid species (introduced)
<i>Parachromis managuense</i> Kullander 1997
<i>Oreochromis aureus</i> Steindachner 1864
<i>Oreochromis niloticus</i> Linnaeus 1758
Lake Xiloá – Midas cichlid species (endemic)
<i>Amphilophus xiloaensis</i> Stauffer and McKaye 2002
<i>Amphilophus amarillo</i> Stauffer and McKaye 2002
<i>Amphilophus sagittae</i> Stauffer and McKaye 2002
<i>Amphilophus</i> sp. “Fat lips” (Stauffer and McKaye 2002, undescribed)
Lake Xiloá – other cichlid species (native)
<i>Astatoheros longimanus</i> Jordan <i>et al.</i> 1930
<i>Archocentrus centrarchus</i> Jordan <i>et al.</i> 1930
<i>Amphilophus rostratus</i> Kullander 1996
<i>Parachromis dovii</i> Kullander <i>et al.</i> 1997
<i>Hypsophrys nicaraguensis</i> Kullander <i>et al.</i> 1997
<i>Parachromis managuense</i> Kullander <i>et al.</i> 1997
<i>Hypsophrys nematopus</i> Chakrabarty <i>et al.</i> 2007
<i>Amantitlania siquia</i> Schmitter-Soto 2007

Supplementary Table 2: Coordinates and length of the transects in lakes Apoyo and Xiloá. Lengths were calculated by measuring start and end coordinates of each transect with a GPS device.

Lake	Transect	Start coordinate	Length [m]
Apoyo	1	11°54,554' N / 86°02,467' W	120
	2	11°54,183' N / 86°01,791' W	115
	3	11°55,626' N / 86°00,854' W	80
	4	11°56,196' N / 86°01,371' W	80
	5	11°56,002' N / 86°03,391' W	80
	6	11°92,538' N / 86°05,557' W	80
Xiloá	1	12°23,120' N / 86°31,857' W	40
	2	12°23,081' N / 86°32,259' W	40
	3	12°21.483' N / 86°32,548' W	50
	4	12°21.428' N / 86°31,510' W	50

Supplementary Table 3: Averaged numbers of cichlid fish per 10 m transect for each transect and depth level. Numbers are the averaged fish counts by the two buddy team partners including the replicates from diving back and forth.

Lake	Transect	Depth [m]				
		5	10	15	20	total
Apoyo	1	-	7.0	6.6	5.7	6.4
	2	12.3	21.7	9.9	13.2	14.3
	3	2.4	23.0	37.3	-	20.9
	4	16.2	21.8	20.4	20.1	19.6
	5	4.3	0.0	0.6	3.2	2.0
	6	1.6	6.9	6.8	7.2	5.6
	total	7.4	14.7	15.0	8.7	11.4
Xiloá	1	43.3	12.6	29.1	9.6	23.7
	2	55.0	5.5	-	37.3	32.6
	3	19.4	20.9	11.2	-	17.2
	4	9.7	3.2	6.4	15.1	8.6
	total	31.9	10.6	15.6	20.7	19.7

Chapter 9

Chapter 9.3

Testing the stages model in the adaptive radiation of
cichlid fishes in East African Lake Tanganyika



Cite this article: Muschick M, Nosil P, Roesti M, Dittmann MT, Harmon L, Salzburger W. 2014 Testing the stages model in the adaptive radiation of cichlid fishes in East African Lake Tanganyika. *Proc. R. Soc. B* **281**: 20140605. <http://dx.doi.org/10.1098/rspb.2014.0605>

Received: 12 March 2014
Accepted: 22 June 2014

Subject Areas:
evolution

Keywords:
trait evolution, convergence, niche overlap, ecological specialization, gill raker

Author for correspondence:
Walter Salzburger
e-mail: walter.salzburger@unibas.ch

One contribution to a Special feature 'Evolution of specialization: insights from phylogenetic analysis'.

Electronic supplementary material is available at <http://dx.doi.org/10.1098/rspb.2014.0605> or via <http://rspb.royalsocietypublishing.org>.

Testing the stages model in the adaptive radiation of cichlid fishes in East African Lake Tanganyika

Moritz Muschick^{1,2}, Patrik Nosil², Marius Roesti¹, Marie Theres Dittmann^{3,4}, Luke Harmon⁵ and Walter Salzburger¹

¹Zoological Institute, University of Basel, Basel 4051, Switzerland

²Department of Animal and Plant Sciences, University of Sheffield, Sheffield S10 2TN, UK

³Institute of Agricultural Sciences, ETH Zurich, Zürich 8092, Switzerland

⁴Clinic for Zoo Animals, Wildlife and Exotic Pets, Vetsuisse Faculty, University of Zurich, Zürich 8057, Switzerland

⁵Department of Biological Sciences and Institute for Bioinformatics and Evolutionary Studies (IBEST), University of Idaho, Moscow, ID, USA

Adaptive radiation (AR) is a key process in the origin of organismal diversity. However, the evolution of trait disparity in connection with ecological specialization is still poorly understood. Available models for vertebrate ARs predict that diversification occurs in the form of temporal stages driven by different selective forces. Here, we investigate the AR of cichlid fishes in East African Lake Tanganyika and use macroevolutionary model fitting to evaluate whether diversification happened in temporal stages. Six trait complexes, for which we also provide evidence of their adaptiveness, are analysed with comparative methods: body shape, pharyngeal jaw shape, gill raker traits, gut length, brain weight and body coloration. Overall, we do not find strong evidence for the 'stages model' of AR. However, our results suggest that trophic traits diversify earlier than traits implicated in macrohabitat adaptation and that sexual communication traits (i.e. coloration) diversify late in the radiation.

1. Introduction

Adaptive radiation (AR) is the rapid diversification of an evolutionary lineage into an array of species as a consequence of their adaptation to various ecological niches and is thought to be responsible for a great deal of the taxonomic, morphological and ecological diversity on Earth [1–3]. ARs are triggered by ecological opportunity through the colonization of novel environments or the evolution of key innovations, opening up new adaptive zones for organisms to specialize and diversify into [1]. Darwin's finches on the Galapagos archipelago [4], anole lizards on the islands of the Caribbean [5] and the species-flock of cichlid fishes in the East African Great Lakes [6] are famous examples of extant ARs. Studying such outbursts of organismal diversification has revealed a number of putatively general features of AR (reviewed in [2]). Among those is the observation that diversification sometimes proceeds more rapidly in the initial phases of an AR ('early burst', EB scenario), but slows down with the filling up of ecological niche space as more species form. However, the generality, or even frequency, of this pattern was recently called into question [7]. Another observation is that the invasion of niche space by diversifying organisms is not random. Different aspects of the environment have been proposed to influence diversification in different phases throughout the course of an AR. Accordingly, there should be temporal stages of AR, in which specialization to available niches, and hence diversification, is predominantly based on different adaptive traits or trait complexes [8].

In vertebrates, for example, it is regularly observed that clades forming early in an AR are subdivided ecologically with respect to macrohabitat specializations [2,8]. Here, we define macrohabitat as a geographically extensive part of the environment encompassing considerable ecological variation, e.g. the benthic or limnetic zones of a lake. This first 'stage' is evident in established ARs, such

as Lake Malawi cichlids [9], and also in very recent, incipient ARs, like crater-lake cichlids [10] or three-spine stickleback [11]. Subsequently, in a second stage, specialization occurs primarily with respect to either more spatially restricted microhabitats or resources therein, as found in e.g. some *Anolis* lizards [5] and Darwin's finches [12]. According to Streebman & Danley's [8] *general vertebrate model*, the most closely related species within an AR often differ in little else than signalling characters, like nuptial coloration. This third and final stage of an AR is observed, for example, in parrotfish [13], Lake Malawi cichlids [9] and elephantfish [14]. It is important to note that these stages are not necessarily discrete and that selection pressures which dominate in one period of a radiation are probably also acting at other times, albeit to a lesser degree [9]. While these hypothesized stages are apparent in some groups, the existence of stages is less clear in other ARs such as Hawaiian drosophilids [15], potentially emphasizing this feature to be unique to vertebrate radiations and definitively calling for further in-depth quantitative evaluations of the 'stages model'.

In vertebrate ARs, it also appears that the sequence of stages relative to another might differ. In *Phylloscopus* Old World leaf warblers [16] or extinct actinopterygians [17], for example, habitat divergence followed trophic divergence. Overall, it is unclear how pervasive the phenomenon of 'AR in stages' actually is in nature. Although theoretical work [18–20] points to the model having merit, empirical testing is hampered by the need to obtain data for a number of traits for many member species of an AR. Synthesizing different studies of trait evolution into a test of the 'stages model' is further complicated by the lack of taxonomic overlap between studies, unclear phylogenetic relationships, the study of different traits, and/or the application of different analytical approaches.

Nonetheless, several predictions can be derived from the hypothesis of 'AR in stages' and tested given the appropriate data: (i) adaptive traits or trait complexes will differ in their amount of phylogenetic signal and time-dependence of their diversification; (ii) the ordering of traits or trait complexes by their time-dependence of diversification should mirror the hypothesized order of stage-wise dominating selection pressures (first: macrohabitat, second: microhabitat and resources, third: sexually selected characters; [8]); and (iii) traits or trait complexes that are involved in specialization to available ecological niches early in an AR should have attained a larger between-species difference, if standardized by the variation within species.

Here, we test these predictions in the AR of cichlid fishes of Lake Tanganyika (LT), East Africa. We investigate the evolution of ecologically and reproductively relevant traits in 51 representative species using phylogenetic comparative methods. Our dataset comprises the majority of cichlid tribes present in LT and a reasonable fraction of the species, including the most abundant ones coexisting in the lake's southern basin. It also covers most of the ecological specializations found in LT cichlids, e.g. epilithic algae grazing, scale eating, fish hunting, invertebrate picking, as well as sand, rock or open water dwelling species. Trait data for body shape, size and weight, lower pharyngeal jaw (LPJ) bone shape and weight, as well as stable isotope data, and a robust multi-marker phylogeny are available from the study of Muschick *et al.* [21]. For this study, we combined these previous data with new data on gut length,

brain weight, gill raker structure and coloration. Trait evolution in LT cichlids has been the focus of previous studies [21–28]. However, most studies considered only one or few traits in isolation and did not comprehensively compare multiple traits in the context of ecology and phylogeny (but see [25]), and the 'stages model' has not yet been tested explicitly in LT cichlids.

In this study, we first test for a phenotype–environment correlation in the traits and in the trait complexes under examination by using stable isotope ratios as a proxy for macrohabitat and trophic niche. Such a phenotype–environment correlation is an inert feature of an AR and informs about the adaptive nature of the traits in question [1]. We then quantify the overlap between species in morphological trait space as proxy for their degree of specialization in the respective trait or trait complex. Correlations among trait complexes that take into account phylogenetic relationships are also examined. Finally, by explicitly fitting models of trait evolution to the trait data and molecular phylogeny, we evaluate the merit of the 'stages model' of AR for LT cichlids.

2. Material and methods

(a) Sampling

We collected trait data for 51 LT cichlid species, which is approximately one-quarter of the endemic species of this lake. The dataset comprises 36 genera (of 53) and 10 of the 16 tribes [29] (see the electronic supplementary material) described for LT cichlids and is thus representative of the phylogenetic, morphological, ecological and behavioural diversity. In subsets of specimens (electronic supplementary material, table S1), we measured six trait 'complexes': body shape, the LPJ apparatus, the gill raker apparatus, brain weight, intestine tract length and colour. The data for body and LPJ shape, stable isotopes and phylogenetic relationships were taken from [21]. Measurements of gill raker traits, gut length, brain weight and scoring for coloration were newly generated for this study.

(b) Choice of traits and their ecological relevance

Body shape is important in swimming performance and manoeuvrability and has been shown to correlate with macrohabitat (i.e. on a benthic–limnetic axis) in cichlids [21]. The LPJ is part of the cichlids' pharyngeal jaw apparatus, i.e. a second set of jaws in the cichlids' throat used to manipulate the food items taken up by the oral jaws [30,31]. Another important component of the feeding apparatus are gill rakers, which are used to filter and sort food items in the buccal cavity in many groups of fishes [32,33–35] (see electronic supplementary material, figure S1). These bony protrusions on the gill arches have been little studied in cichlids, as opposed to other evolutionary model systems such as stickleback [32]. After uptake, mastication and filtering, food items reach the intestinal tract where enzymatic digestion takes place and nutrients are absorbed. Herbivorous species specialized on resources of low digestibility, e.g. algae and plants, usually have longer intestines resulting in a longer retention time for improved digestion [27]. Brain size is known to show a strong allometric relationship with body size over a large range of organisms [36], but residual variation and shifts in relative sizes of brain parts have been hypothesized to have adaptive value [37,38]. Body coloration in cichlids can differ greatly between species, even between closely related ones or populations of the same species [39], and is important not only in mate recognition, reproductive behaviour, intraspecific aggression, but also camouflage and mimicry [40].

(c) Trait data

The gill raker trait assessments essentially followed previous investigations in three-spine stickleback [32,34]. Brain tissue was removed from the neurocranium in the field and stored in ethanol or RNAlater; in the laboratory, preserved fish brains were drained and dried at 60°C overnight and subsequently weighed to the nearest milligram. To investigate gut length, we removed the entire alimentary canal ('gut') from the anus to the posterior end of the stomach and measured its length to the nearest millimetre. In order to evaluate body coloration, we adopted and modified an existing colour-scoring scheme for cichlids [41].

In the following, we describe data re-used from Muschick *et al.* [21]: phylogenetic relationships were derived from the enforced molecular-clock phylogeny by pruning it to the 51 species included here. Body shape information was assessed on the basis of landmarks derived from photographs using TPSDIG [42], Procrustes aligned in MORPHO [43] and analysed in R [44]; LPJ shape information was obtained in a similar way, but including a sliding process of semi-landmarks in TPSRELW [45]. Stable isotope data were used as proxies for macrohabitat-related specialization on a benthic–limnetic axis ($\delta_{13}\text{C}$) and for specialization to the trophic niche ('microhabitat', $\delta_{15}\text{N}$; e.g. [46]).

Prior to statistical analyses, we log transformed all trait values, apart from landmark Procrustes coordinates, gill raker counts and coloration scores, using the `phyl.resid` function of the R package PHYTOOLS [47]. Further details to trait data and their statistical treatment are provided in the electronic supplementary material.

(d) Correlations between traits and the ecological niche

As a test for ecological specialization, we evaluated the correlation of trait values with ecological niches, where stable isotope ratios of carbon and nitrogen were used as niche proxies. The relationship of body and LPJ shape and gut length with stable isotope ratios has been investigated before [21,27], but for gill raker morphology in cichlids this is the first demonstration to our knowledge. We correlated the first principal component (PC) of a PC analysis (PCA) comprising both stable isotope ratios, as well as each element separately, with the first PC for each trait (respective scaled trait values for univariate traits, neither corrected for phylogenetic relationships) and accounted for phylogenetic relationships between species using phylogenetic generalized linear models as implemented in the R package CAPER [48].

(e) Ecological specialization and overlap between species

We used plots of linear discriminants (LDs) to illustrate each species' position in morphospace for the multivariate data, and boxplots to illustrate the univariate data. Next, we calculated the between-species distances for each trait (see the electronic supplementary material for details). To compare the relative overall separation of species for each trait, we used 'species' as the independent variable in multivariate analyses of variance (MANOVA) for multivariate traits and in analyses of variance (ANOVA) for univariate traits, respectively. Here, we could not include colour as a trait, as no within-species measurements were available. We used Wilks' λ to assess species overlap in the MANOVA, and F -values in the ANOVA. As Mahalanobis distances are scaled by the within-group variance, we used them as a generalized measure of trait divergence, which can be compared among traits. To reveal the ordering of traits by their attained trait distance, we implemented breakpoint regression models following [14]. We estimated breakpoints and respective linear relationships for segments using the function `segmented` in the SEGMENTED R package [49]. For multivariate traits, we calculated the

Mahalanobis distances as described above. Univariate trait values were scaled with the averaged within-species variance.

(f) Phylogenetic tests for stages of adaptive radiation

To test for apparent stages in diversification, we fitted models of trait evolution to the trait axes derived from our transformation of raw trait values (see above) using the `fitContinuous` function in the R package GEIGER [50] (number of random starting points = 1000, simultaneous estimation of standard error). To describe the more general process of trait evolution, we fitted three macroevolutionary models to our data: (i) the Brownian motion model assumes trait values to evolve according to a diffusion process, resulting in trait similarity between species being mainly dependent on the amount of shared ancestry; (ii) the Ornstein–Uhlenbeck model of trait evolution simulates attraction to a single optimum of trait values, with the alpha parameter indicating the strength of this attraction; and (iii) the white noise model does not assume a covariance structure in the data owing to phylogeny, it is equivalent to drawing trait values from a single normal distribution. We also assessed the time-dependence of trait evolution using two models. The δ model by Pagel [51] was used to assess the relative contribution of trait evolution early in the radiation versus late in the radiation. Here, values below 1 indicate that trait evolution occurred primarily along the more basal branches in the phylogeny, whereas values greater than 1 indicate trait evolution predominantly in younger subclades. Second, the a parameter of the EB model (also known as ACDC, for 'accelerating–decelerating' [52]) implemented in GEIGER was used to test for accelerating or decelerating rates of trait evolution across the phylogeny. Negative values indicate a slowdown in trait evolution, while positive values identify acceleration. For an estimate of phylogenetic signal in the data, we calculated Pagel's λ [51], where λ ranges from 0 to 1 and higher λ values mean stronger phylogenetic signal in the trait data. As a second measure of the pervasiveness of phylogenetic signal in our trait data, we calculated Blomberg's K statistic [52,53] using the `phylosignal` function in the R package PICANTE v. 1.6–1 [54]; $K > 1$ indicates stronger resemblance of species than expected under a Brownian model of trait evolution, while values less than 1 point to a greater evolutionary malleability of the trait.

If the diversification of LT cichlids was indeed driven by different selection pressures in the hypothesized order of stages, we would expect this to be reflected in parameter estimates and the fit of the macroevolutionary models to the data. The time-dependent models should fit reasonably well and differ in their parameter estimates between traits, pointing to different temporal maxima of divergence. To robustly deduct a temporal dependence of diversification, however, the Ornstein–Uhlenbeck model of trait evolution would need to be rejected [7], as selection to a single trait optimum with differing strength could mimic such time-dependence.

3. Results

(a) Correlations between traits and the ecological niche

The ecological relevance of traits can be assessed by their correlation with parameters describing the ecological niche. Here, we used stable isotope ratios of carbon (C) and nitrogen (N) as proxies for macrohabitat preference and trophic position of species, respectively (figure 1 and table 1). Body shape is significantly correlated with $\delta_{13}\text{C}$, but not with $\delta_{15}\text{N}$. The correlation with $\delta_{13}\text{C}$ becomes weaker and insignificant when phylogenetic relationships are taken into account (by phylogenetic generalized least squares (PGLS)). LPJ shape and gill raker morphology correlate significantly with both

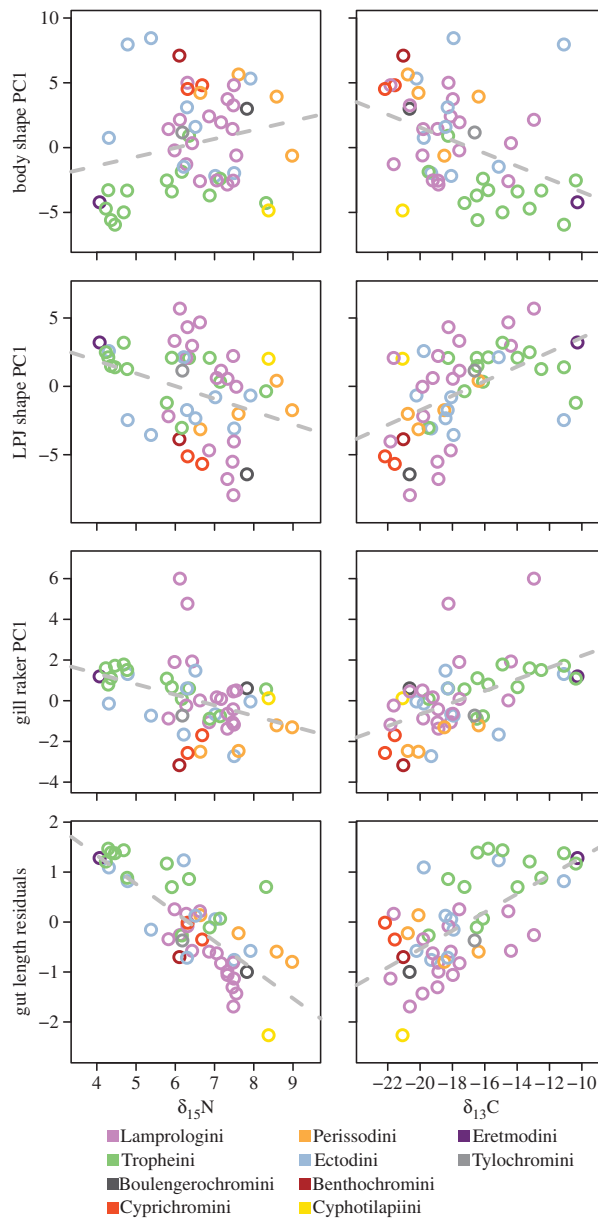


Figure 1. Phenotype–environment correlations. Species means of trait values (see ‘Material and methods’ for details) were plotted against species means of $\delta_{13}\text{C}$ and $\delta_{15}\text{N}$. $\delta_{13}\text{C}$ is considered to be an indicator of macrohabitat use, with higher or lower values reflecting a benthic or limnetic carbon source, respectively. $\delta_{15}\text{N}$ is a proxy for trophic level, with larger values reflecting a higher trophic position.

isotopic signatures, but more strongly with $\delta_{13}\text{C}$. Phylogenetic correction decreases these correlations in LPJ for both isotopes and in gill rakers for $\delta_{15}\text{N}$, while it reinforces the correlation between $\delta_{13}\text{C}$ and gill rakers. Gut length correlates very strongly with both isotopic signatures. Here, both correlations are even more pronounced in the PGLS analysis. Brain weight and coloration stand in stark contrast to the aforementioned traits in that they either do not correlate with ecological niche proxies, or, in the case of $\delta_{13}\text{C}$ and colour, only with marginal significance.

Between traits, partial Mantel tests using phylogenetic distance as a covariate revealed weak to moderate and predominantly positive relationships (table 2). However, after correction for multiple comparisons, only a subset of the correlations remained statistically significant. The strongest correlations were found between gut length and brain weight (partial Mantel statistic = 0.5, $p < 0.0001$), and between gill raker traits and LPJ traits (0.42, $p < 0.0001$). Gut length and LPJ shape correlated positive with a coefficient of 0.24. No correlation was evident between colour and any other trait.

Table 1. Phenotype–environment correlation. (Correlations of species means for the first principal component of residual trait data (phylogeny not accounted for) and the stable isotope ratio of either carbon or nitrogen (Pearson's correlation coefficient). Phylogenetically generalized least-squares (PGLS) analysis was used to remove the phylogenetic signal potentially present in this correlation. Correlations that remained significant after table-wide adjustment of p -values (after Bonferroni) are given in *italics*.)

		body shape	LPJ shape	gill raker	gut length	brain weight	colour
$\delta_{13}\text{C}$	Pearson's correlation	−0.41	0.51	0.53	0.63	0.12	0.28
	p -value	0.0031	<i>0.0002</i>	<i>0.0001</i>	<i>0.0000</i>	0.3878	0.0497
	PGLS correlation	−0.12	0.38	0.72	2.03	0.02	0.59
	p -value	0.3752	0.0278	0.0203	<i>0.0011</i>	0.9817	0.0591
$\delta_{15}\text{N}$	Pearson's correlation	0.22	−0.35	−0.37	−0.77	−0.22	−0.18
	p -value	0.1197	0.0137	0.0076	<i>0.0000</i>	0.1216	0.2165
	PGLS correlation	−0.06	−0.12	−0.31	−1.06	−0.21	0.01
	p -value	0.1940	0.0563	0.0040	<i>0.0000</i>	0.3718	0.9611

Table 2. Trait complex covariation among Tanganyikan cichlids. (Results of partial Mantel tests accounting for phylogenetic distance. Mahalanobis distances were calculated for traits with intraspecific variance, and Manhattan distance was calculated for body coloration. Correlations are given below the diagonal, p -values above. Comparisons that are significant after adjustment for multiple comparisons (after Bonferroni) are given in *italics*.)

Mantel statistic	p -value					
	body shape	LPJ shape	gill raker	gut length	brain weight	colour
body shape		<i>0.0004</i>	0.0044	0.0035	0.0286	0.595
LPJ shape	0.22		<i>0.0001</i>	<i>0.0002</i>	<i>0.0015</i>	0.5824
gill raker	0.15	0.42		0.1744	<i>0.0027</i>	0.7944
gut length	0.15	0.24	0.05		<i>0.0001</i>	0.2837
brain weight	0.11	0.19	0.18	0.5		0.7686
colour	−0.02	−0.02	−0.06	0.03	−0.06	

(b) Ecological specialization and overlap between species

The LT cichlid species examined are at least somewhat, and often strongly, separated in the traits studied here (figure 2). The axes along which species are most separated (scaled by within-species variance) were highlighted by a LD analysis. In body shape space (figure 2a), species show fewer areas of between-species overlap than in the other two multivariate traits, LPJ (figure 2b) and gill rakers (figure 2c). The between-species overlap is most apparent in gill raker space. Gut length (figure 2d) and brain weight (figure 2e) also separate species in morphospace, with overlap between species being less pronounced in gut length. The 14 algae-eating species in the dataset clearly show increased relative gut lengths compared with the non-algae-eating species. Overlap in colour space could not be examined, but some clustering is apparent (figure 2f). On the tribe-level, traits also do show separation, with partitioning of morphospaces being most apparent between the most species-rich LT tribes Lamprologini, Tropheini and Ectodini.

MANOVA and ANOVA show that means of trait axes are significantly different between species for all traits (electronic supplementary material, table S2). Wilks' λ indicates that almost all the variance in the multivariate traits is accounted for by species identity. Compared with body and LPJ shape, gill rakers show increased, but still minor, unexplained

variance, interpretable as niche overlap (Wilks' $\lambda = 0.0121$). The lower F -value in the analysis of relative brain weight compared to gut length corroborates the larger overlap between species in this trait (electronic supplementary material, table S2). In summary, the collective results indicate that the species are well differentiated in all traits, but most strongly in body shape.

(c) Trait evolution and test for the 'stages model'

Recall that the 'stages model' of AR predicts that macrohabitat-related traits, such as body shape, diversify early in the radiation (stage 1). Successively, trophic traits (stage 2) and then traits involved in sexual communication (stage 3) would come to dominate diversification. By fitting models of trait evolution to our data, we evaluated the plausibility of a scenario of diversification in stages in LT cichlids. We assessed which macroevolutionary model fits the data best by comparing the sample-size corrected Akaike information criterion (AIC_c). The different body shape dimensions derived from the phylogenetic PCA (pPCA) show the best fit to different models of trait evolution: body shape dimension 1 is best approximated by the Pagel's λ model, whereas dimensions 2 and 3 are best approximated by Ornstein–Uhlenbeck and Pagel's δ models, respectively. However, while the model fit difference for dimension 1 is moderately pronounced (min. $\Delta\text{AIC}_c = 1.47$), the model selection is ambivalent in the other body shape

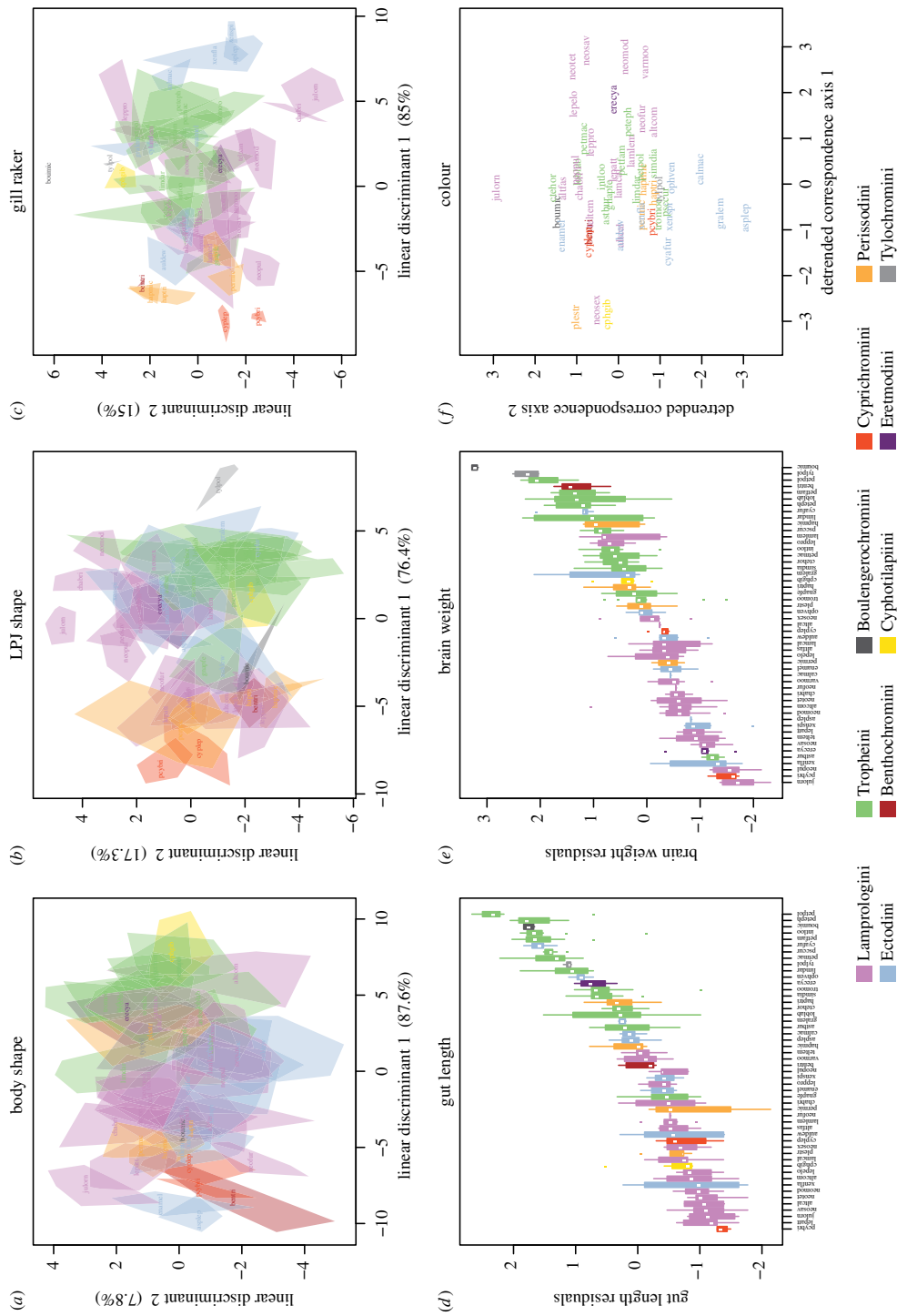


Figure 2. Niche overlap and divergence between species of cichlids. (a–c) Species overlap on the first two axes from a LD analysis for each multivariate trait. Polygons for species are defined by the most extreme individual and coloured by tribe identity. Amount of between-group variance explained by each axis is given in parentheses. (d,e) Species overlap for the univariate traits gut length and brain weight, sorted by median. (f) Relative position of species in colour space as defined by the first two axes of a detrended correspondence analysis.

dimensions. This is similar for LPJ shape, where the min. ΔAIC_c is 0.47 for dimension 1 and even lower for dimensions 2 and 3. Gill raker traits, gut length and brain weight all show a phylogenetic signal very similar to that produced by a Brownian motion-like trait evolution (electronic supplementary material, table S3), evidenced by Blomberg's K and Pagel's λ parameter being close to 1, and best fit of the Brownian motion model of trait evolution. Coloration appears to be best fitted by a model of Brownian trait evolution, too. However, in contrast to most other traits, coloration dimensions are reasonably well fitted by the 'white noise' model, as well. This might indicate a lack of phylogenetic covariance structure in this trait. We used Pagel's δ model and the EB model to reveal variation in the rate of trait evolution across the timeframe of the radiation. All traits show an acceleration in trait evolution as evidenced by positive a values, as well as a concentration of trait evolution comparatively early in the radiation, shown by δ values more than 1 (electronic supplementary material, table S3). Among the six trait complexes, colour appears to be most rapidly evolving. Of the remaining traits, body shape evolution appears to accelerate in the course of the radiation, followed by LPJ shape. Gill rakers, gut length and brain weight show comparatively little acceleration. However, in no case was the fit of a time-dependent model significantly better than either the Brownian motion or the Ornstein–Uhlenbeck model of trait evolution ($\Delta AIC_c > 2$). Thus, we have to limit our tentative conclusions about time dependency of trait evolution to the above-stated fits of macroevolutionary scenarios with our data.

In quantifying phylogenetic signal, our results are in agreement with those of Wagner *et al.* [27], who reported a Pagel's λ of 0.995 for gut length, and of Gonzalez-Voyer *et al.* [24] who reported a Pagel's λ of 0.71 for brain weight. Both studies have investigated LT cichlids, but included less and different species than this study. Clabaut *et al.* [22] find that body shape is best predicted by trophic niche and water depth, which is corroborated by our analyses here.

Breakpoint regression models for distances of residual trait values (phylogeny not being accounted for) scaled by within-species variance against phylogenetic distance show the initial increase in trait divergence with evolutionary divergence. After the estimated breakpoint, linear models indicate no further increase in trait divergence, and trait distance appears to be uncorrelated to phylogenetic distance (electronic supplementary material, figure S3).

4. Discussion

Our examination of the stages model of AR in LT cichlids is based on data for six trait complexes analysed within a phylogenetic framework using 51 representative species. We also examined the ecological relevance of these traits using stable isotope data, specialization to ecological niches and niche overlap, as well as trait covariation.

(a) Phenotype–environment correlation

All traits apart from brain size and coloration showed correlation with the ecological niche as approximated by stable isotope data. Overall residual brain size as measured in this study does not appear to correlate with ecological niche. However, it seems likely that cognitive demands regarding specialization to macrohabitats and trophic niches differ such that an effect on brain evolution would be expected

[55]. For a smaller dataset of LT cichlids, a significant correlation between diet and brain weight was found [56], but their approximation of the ecological niche differs considerably from ours as it relied on qualitative descriptions from the literature. In general, specialization to available niches might involve changes in the relative sizes of brain substructures [57,58] such that the measure of whole brain weight is too crude to characterize differences.

In contrast to brain weight, gut length is strongly correlated with the trophic niche, and also with our proxy for the macrohabitat niche, $\delta_{13}C$. This is not surprising given the demands of a herbivore's diet on the digestive system. Herbivores are also more likely to acquire their carbon signature from the littoral realm, as this is where epilithic algae occur, explaining the correlation with our macrohabitat proxy. An example of an exception to this is *Cyathopharynx furcifer* (Cyafur), a plankton feeder with a limnetic carbon signature. Figure 2e illustrates this clearly with herbivorous species being those with the highest relative gut length. The gill raker data also meet the expectation of a significant correlation with ecological niche. Coloration does correlate with marginal significance with the macrohabitat niche. The colour dimorphism found in some species has been correlated with microhabitat specializations [59,60], so this pattern could also be expected on the species level. However, convergence through mimicry might obscure patterns [61,62]. More precise quantification of colour and its relationship to niche specialization is thus required.

(b) Ecological specialization and overlap between species

By scaling trait values by the average within-species variance, we standardized our measurements to allow comparisons across traits and to infer the relative contribution to niche specialization. We find that traits varied in their overlap between species. While body shape shows signs of a more recent divergence compared with trophic traits, it is the trait attaining the largest between-species trait distances if scaled by within-species variance. Also, species overlap less in body shape morphospace compared to trophic traits. This suggests a higher degree of specialization of species in their macrohabitat niche than in trophic characters, such as the pharyngeal jaw apparatus. Trophic traits attain less between-species standardized trait distance, which could be interpreted as trophic traits being less evolutionarily constrained. Another possible explanation is that the adaptive landscape of microhabitats differs between geographically separated sites, which might inflate the morphospace taken up by a given species if individuals from different sites are analysed together.

(c) Trait covariation

Analysing trait covariation while controlling for phylogenetic relatedness reveals several pairwise comparisons to be significantly correlated. Gut length and brain weight covary—most probably due to both being trophic adaptations, for example to herbivory, which could also impose specific cognitive demands due to habitat complexity. The weaker correlation of body and LPJ shape is somewhat unexpected, as body shape is thought of conferring adaptation to macrohabitat, whereas the pharyngeal jaw apparatus is mainly involved in

trophic adaptation. However, food types and abundances differ not only between microhabitats but also between macrohabitats, and, thereby, macrohabitat and trophic adaptations might covary. The correlation between LPJ shape and gill raker traits is not surprising, given that both traits together are responsible for the uptake and the processing of food items, possibly leading to functional constraints. In addition, developmental constraints may further limit independency of trait evolution in the case of LPJ and gill rakers, as both trait complexes derive from gill arch constituents.

(d) Trait evolution and test for the 'stages model'

The predictions derived from the 'stages model' were met to different degrees. In general, we did not find definitive evidence for the existence of discrete stages in the AR of cichlids in LT, because the time-dependent models of phenotypic evolution (EB model and Pagel's δ) were not supported over others. However, some of the patterns observed were nonetheless consistent with the 'stages model'. For example, traits were found to exhibit different amounts of phylogenetic signal and the timing of diversification varied among traits (supporting our prediction (i)). The ordering of trait diversification (prediction (ii)), with the caveat of not being able to reject the Ornstein–Uhlenbeck model, did not mirror our expectations in the amount of phylogenetic signal. The ordering of the relative amounts of attained relative between-species divergence (scaled by within-species divergence) is in agreement with our prediction from the 'stages model' and the findings of the phylogenetic model fitting.

The time-dependence of traits analysed here was not unequivocally evident. Simpler models, not invoking a change in the rate of phenotypic evolution within the timeframe of the AR, received greater support. That said, the time-dependent models (EB model and Pagel's δ) often did fit reasonably well even compared to the respective best-fit model ($\Delta AIC_c < 2$). Parameter estimates for those two models differed between traits, but did not conform to the expectation of an apparent ordering of trait evolution in the form of macrohabitat-related traits first, then microhabitat-related (in our case: trophic) traits, and, finally, sexual communication traits. Instead, our parameter estimates revealed a different, tentative ordering of trait evolution in LT cichlids that was sometimes but not unambiguously corroborated by the fit of the respective macroevolutionary models: it appears that the evolution of trophic traits is less accelerated across the radiation than either body shape or coloration, suggesting that there are temporal stages of phenotypic evolution in the AR. The ordering, however, is somewhat different to the 'stages model', with trophic traits diverging first, followed by macrohabitat-related traits, and, finally communication traits.

The same conclusion can be drawn from the comparison of the amount of phylogenetic signal in each trait complex. Traits characterized by their functionality in feeding (LPJ, gill raker, gut) showed a stronger phylogenetic signal than traits used in macrohabitat adaptation (body shape) or communication (colour). In fact, coloration showed no significant phylogenetic signal at all, which is in agreement with the stages model [8] and is to be expected given current arguments that the evolution of sexual characters is often rapid and unconstrained. On the other hand, the temporal patterns of phenotypic evolution in trophic relative to macrohabitat-related traits do not conform to the 'stages model'. All results of our model fitting approach point to less phylogenetic signal and, hence, a more

recent divergence in body shape traits compared with trophic traits. A similar scenario, with diversification into macrohabitats coming second, has been suggested for *Phylloscopus* warblers [16], and, more recently, for two extinct fish radiations [17]. The latter authors presented a 'head-first' scenario, in which fish head morphology consistently diversified prior to trunk morphology. Assuming head morphology to be a predominantly trophic trait, and trunk morphology responding mainly to selection pressures exerted by macrohabitats, the order of trait evolution in those radiations appears to follow the same trend as we uncover here. Note, however, that such trait complexes are probably not independent. Body shape, for example, also includes information on head shape, so that a signal of adaptation to diet in head shape is probably captured by body shape data too. Thus, the signal in body shape is perhaps biased towards a higher similarity with trophic traits, a possibly general problem in analyses like this. An alternative interpretation of our results is that coloration was a target of selection throughout the radiation, whereas body shape and, especially, the trophic trait complexes LPJ, gill rakers and gut length were involved in certain (initial) stages only [63].

Yet another explanation for our findings is that convergent evolution within the AR of LT cichlids [21] caused an 'erosion' of the phylogenetic signal in certain trait complexes, possibly accumulating with time since the climax of divergence. Convergence might also be a potential explanation of why divergence with respect to macrohabitat was suggested as the initial stage of morphological evolution in the first place [8,9]. Low within-species but large between-species variance in body shape make this trait appear well suited for taxonomic inference, but a lack of phylogenetic signal would lead to erroneous taxonomic groupings. In cases where convergent evolution has in this way hampered the traditional reconstruction of phylogenetic relationships that did not include molecular data, stages of AR could have been suggested spuriously. By relying on taxonomic affiliations derived from characters implicated in, for example, habitat or resource specializations, recurrent adaptations would not be recognized as such. In this case, the impression of a temporal order of phenotypic evolution within ARs would be an artefact, as has been shown in bower-building cichlids of Lake Malawi [64]: molecular phylogenetic analyses revealed that some genera were actually not monophyletic to the result that the revised trait distribution suggested trophic morphology to diverge consistently earlier in the AR compared to mate recognition traits (i.e. bower shape). Similarly, in LT cichlids such systematic revisions have been common as reliable molecular phylogenies have become available that uncovered cases of convergent evolution (e.g. [21,65]).

(e) Limitations of our approach

The inability of our study to conclusively reject or support a model of AR in stages in LT cichlids has several plausible explanations, including study design, methodological limitations or biological processes. Concerning study design, the basal tribes Trematocarini and Bathybatini, both predominantly inhabiting the open waters, are under-represented in our dataset. Including more species of those tribes might help generate more definitive results with respect to the 'stages model'.

Methodologically, our approach analysed body shape as a whole, not discriminating between head- and trunk shapes.

Therefore, trophic adaptations in head shape could be represented in our assessment of body shape evolution (see above). However, Muschick *et al.* [21] have shown that the major axis of body shape evolution in LT cichlids discriminates deep-bodied versus elongated morphs, reflecting macrohabitat adaptation also in other fish groups [10,66]. Additionally, although our dataset is certainly rich relative to today's standards and can readily address questions about ecological specialization and the order of divergence of traits early in the radiation, the number of taxa used might be insufficient to reliably discriminate between scenarios of recent trait evolution (e.g. [67]). Also in terms of methodology, the fitting of evolutionary models implicitly uses reconstructed ancestral phenotypes, which may be inaccurate. Additional information from fossils would be highly useful to verify these estimates [68]. Furthermore, evolutionary change might be underestimated in cases where later changes curb earlier ones, to the end that differences in the rate of evolution between traits might become blurred [69].

5. Conclusion

In this study, we examined the time-dependence of trait evolution and diversification in the species-flock of cichlid fishes in East African LT to test whether this AR proceeded in discrete stages, as has been proposed earlier for vertebrate ARs. Although we do not find strong evidence for the classic stages model of AR in LT cichlids, we find that—contrary to earlier predictions—trophic traits diversified earlier in the

radiation than traits related to macrohabitat specializations, whereas sexual communication traits (i.e. coloration) appear to have diversified late. The lack of power in our approach to discriminate between plausible macroevolutionary hypotheses emphasizes the need for even more comprehensive comparative studies, which would benefit from the addition of fossil data.

Ethics statement. All experiments have been performed under permits issued by the cantonal veterinary office in Basel and the Lake Tanganyika Research Unit, Department of Fisheries, Zambia.

Data accessibility. The datasets supporting this article have been uploaded as part of the electronic supplementary material.

Acknowledgements. We thank the organizers of the special issue, J. Vamosi, S. Renner and S. Armbruster, for inviting us to participate; the associate editor J. Vamosi, R. Gillespie and an additional anonymous referee for valuable comments; our collaborators at the Department of Fisheries, Republic of Zambia, for support in the field and research permits; and H. H. Büscher and various other members of the Salzburger laboratory for help during fieldwork and comments on the manuscript.

Funding statement. M.M. received funding from the Swiss National Science Foundation (SNSF) and the University of Basel, and a travel stipend from the European Science Foundation's (ESF) 'Frontiers of Speciation Research' programme. P.N. and W.S. were both supported by the European Research Council (ERC, grants NatHisGen as well as INTERGENADAPT and CICHlid~X, respectively). W.S. received further support through the University of Basel and the Swiss National Science Foundation (grant nos. 3100A0_138224 and CRSII3_136293). M.R. was supported by a Swiss National Science Foundation (SNSF) Sinergia grant (CRSII3_136293) awarded to W.S. and a Swiss Academy of Sciences (SCNAT) travel grant. Parts of the samples have been collected on an expedition to LT supported by the National Geographic Society.

References

- Schluter D. 2000 *The ecology of adaptive radiation*. New York, NY: Oxford University Press.
- Gavrilets S, Losos JB. 2009 Adaptive radiation: contrasting theory with data. *Science* **323**, 732–737. (doi:10.1126/science.1157966)
- Simpson GG. 1953 *The major features of evolution*. New York, NY: Columbia University Press.
- Lack D. 1947 *Darwin's finches*. Cambridge, UK: Cambridge University Press. (S.I.)
- Losos JB. 2009 *Lizards in an evolutionary tree: ecology and adaptive radiation of anoles*. Berkeley, CA: University of California Press.
- Fryer G, Iles TD. 1972 *The cichlid fishes of the great lakes of Africa: their biology and evolution*. Edinburgh, UK: Oliver and Boyd.
- Harmon LJ *et al.* 2010 Early bursts of body size and shape evolution are rare in comparative data. *Evolution* **64**, 2385–2396. (doi:10.1111/j.1558-5646.2010.01025.x)
- Streelman JT, Danley PD. 2003 The stages of vertebrate evolutionary radiation. *Trends Ecol. Evol.* **18**, 126–131. (doi:10.1016/S0169-5347(02)00036-8)
- Danley PD, Kocher TD. 2001 Speciation in rapidly diverging systems: lessons from Lake Malawi. *Mol. Ecol.* **10**, 1075–1086. (doi:10.1046/j.1365-294X.2001.01283.x)
- Barluenga M, Stöting KN, Salzburger W, Muschick M, Meyer A. 2006 Sympatric speciation in Nicaraguan crater lake cichlid fish. *Nature* **439**, 719–723. (doi:10.1038/nature04325)
- Nagel L, Schluter D. 1998 Body size, natural selection, and speciation in sticklebacks. *Evolution* **52**, 209–218. (doi:10.2307/2410936)
- Grant PR, Grant BR. 2007 *How and why species multiply: the radiation of Darwin's finches*. Princeton, NJ: Princeton University Press.
- Streelman JT, Alfaro M, Westneat MW, Bellwood DR, Karl SA. 2002 Evolutionary history of the parrotfishes: biogeography, ecomorphology, and comparative diversity. *Evolution* **56**, 961–971. (doi:10.1111/j.0014-3820.2002.tb01408.x)
- Amegard ME, McIntyre PB, Harmon LJ, Zelditch ML, Crampton WG, Davis JK, Sullivan JP, Lavoue S, Hopkins CD. 2010 Sexual signal evolution outpaces ecological divergence during electric fish species radiation. *Am. Nat.* **176**, 335–356. (doi:10.1086/655221)
- Kambysellis MP, Ho K-F, Craddock EM, Piano F, Parisi M, Cohen J. 1995 Pattern of ecological shifts in the diversification of Hawaiian *Drosophila* inferred from a molecular phylogeny. *Curr. Biol.* **5**, 1129–1139. (doi:10.1016/S0960-9822(95)00229-6)
- Richman AD, Price T. 1992 Evolution of ecological differences in the Old-World leaf warblers. *Nature* **355**, 817–821. (doi:10.1038/355817a0)
- Sallan LC, Friedman M. 2012 Heads or tails: staged diversification in vertebrate evolutionary radiations. *Proc. R. Soc. B* **279**, 2025–2032. (doi:10.1098/rspb.2011.2454)
- Gavrilets S, Vose A. 2005 Dynamic patterns of adaptive radiation. *Proc. Natl Acad. Sci. USA* **102**, 18 040–18 045. (doi:10.1073/pnas.0506330102)
- Gavrilets S, Vose A. 2009 Dynamic patterns of adaptive radiation: evolution of mating preferences. In *Speciation and patterns of diversity* (eds R Butlin, J Bridle, D Schluter), pp. 102–126. Cambridge, UK: Cambridge University Press.
- Gavrilets S. 2004 *Fitness landscapes and the origin of species*. Princeton, NJ: Princeton University Press.
- Muschick M, Indermaur A, Salzburger W. 2012 Convergent evolution within an adaptive radiation of cichlid fishes. *Curr. Biol.* **22**, 2362–2368. (doi:10.1016/j.cub.2012.10.048)
- Clabaut C, Bunje PM, Salzburger W, Meyer A. 2007 Geometric morphometric analyses provide evidence for the adaptive character of the Tanganyikan cichlid fish radiations. *Evolution* **61**, 560–578. (doi:10.1111/j.1558-5646.2007.00045.x)
- Hoerner ME. 2011 Testing for differences in rates of speciation, extinction, and morphological evolution in four tribes of cichlids endemic to Lake Tanganyika, East Africa. *Evolution* **65**, 3398–3412. (doi:10.1111/j.1558-5646.2011.01390.x)

24. Gonzalez-Voyer A, Winberg S, Kolm N. 2009 Distinct evolutionary patterns of brain and body size during adaptive radiation. *Evolution* **63**, 2266–2274. (doi:10.1111/j.1558-5646.2009.00705.x)
25. Gonzalez-Voyer A, Kolm N. 2011 Rates of phenotypic evolution of ecological characters and sexual traits during the Tanganyikan cichlid adaptive radiation. *J. Evol. Biol.* **24**, 2378–2388. (doi:10.1111/j.1420-9101.2011.02365.x)
26. Duponchelle F, Paradis E, Ribbink AJ, Turner GF. 2008 Parallel life history evolution in mouthbrooding cichlids from the African Great Lakes. *Proc. Natl Acad. Sci. USA* **105**, 15 475–15 480. (doi:10.1073/pnas.0802343105)
27. Wagner CE, McIntyre PB, Buels KS, Gilbert DM, Michel E. 2009 Diet predicts intestine length in Lake Tanganyika's cichlid fishes. *Funct. Ecol.* **23**, 1122–1131. (doi:10.1111/j.1365-2435.2009.01589.x)
28. Kidd MR, Duftner N, Koblmüller S, Sturmbauer C, Hofmann HA. 2012 Repeated parallel evolution of parental care strategies within *Xenotilapia*, a genus of cichlid fishes from Lake Tanganyika. *PLoS ONE* **7**, e31236. (doi:10.1371/journal.pone.0031236)
29. Koblmüller S, Sefc KM, Sturmbauer C. 2008 The Lake Tanganyika cichlid species assemblage: recent advances in molecular phylogenetics. *Hydrobiologia* **615**, 5–20. (doi:10.1007/s10750-008-9552-4)
30. Liem KF, Osse JWM. 1975 Biological versatility, evolution, and food resource exploitation in African cichlid fishes. *Am. Zool.* **15**, 427–454. (doi:10.1093/icb/15.2.427)
31. Liem KF. 1973 Evolutionary strategies and morphological innovations: cichlid pharyngeal jaws. *Syst. Zool.* **22**, 425–441. (doi:10.2307/2412950)
32. Berner D, Adams DC, Grandchamp AC, Hendry AP. 2008 Natural selection drives patterns of lake-stream divergence in stickleback foraging morphology. *J. Evol. Biol.* **21**, 1653–1665. (doi:10.1111/j.1420-9101.2008.01583.x)
33. Smith TB, Skulason S. 1996 Evolutionary significance of resource polymorphism in fishes, amphibians and birds. *Annu. Rev. Ecol. Syst.* **27**, 111–133. (doi:10.1146/annurev.ecolsys.27.1.111)
34. Gerking SD. 1994 *Feeding ecology of fish*. San Diego, CA: Academic Press.
35. Sanderson SL, Cheer AY, Goodrich JS, Graziano JD, Callan WT. 2001 Crossflow filtration in suspension-feeding fishes. *Nature* **412**, 439–441. (doi:10.1038/35086574)
36. Striedter GF. 2005 *Principles of brain evolution*. Sunderland, MA: Sinauer Associates.
37. Barton RA, Harvey PH. 2000 Mosaic evolution of brain structure in mammals. *Nature* **405**, 1055–1058. (doi:10.1038/35016580)
38. Sol D, Bacher S, Reader SM, Lefebvre L. 2008 Brain size predicts the success of mammal species introduced into novel environments. *Am. Nat.* **172**, S63–S71. (doi:10.1086/588304)
39. Baric S, Salzburger W, Sturmbauer C. 2003 Phylogeography and evolution of the Tanganyikan cichlid genus *Tropheus* based upon mitochondrial DNA sequences. *J. Mol. Evol.* **56**, 54–68. (doi:10.1007/s00239-002-2380-7)
40. Maan ME, Sefc KM. 2013 Colour variation in cichlid fish: developmental mechanisms, selective pressures and evolutionary consequences. *Semin. Cell Dev. Biol.* **24**, 516–528. (doi:10.1016/j.semcdb.2013.05.003)
41. Salzburger W, Niederstätter H, Brandstätter A, Berger B, Parson S, Snoeks J, Sturmbauer C. 2006 Colour-assortative mating among populations of *Tropheus moorii*, a cichlid fish from Lake Tanganyika, East Africa. *Proc. R. Soc. B* **273**, 257–266. (doi:10.1098/rspb.2005.3321)
42. Rohlf FJ. 2008 *tpsDig*, v. 2.11. Stony Brook, NY: Department of Ecology and Evolution, SUNY.
43. Klingenberg CP. 2011 Morpho: an integrated software package for geometric morphometrics. *Mol. Ecol. Resour.* **11**, 353–357. (doi:10.1111/j.1755-0998.2010.02924.x)
44. R Core Team. 2012 *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.
45. Rohlf FJ. 2010 *tpsRelw*, v. 1.49. Stony Brook, NY: Department of Ecology and Evolution, SUNY.
46. Vander Zanden MJ, Vadeboncoeur Y. 2002 Fishes as integrators of benthic and pelagic food webs in lakes. *Ecology* **83**, 2152–2161. (doi:10.1890/0012-9658(2002)083[2152:FAIOBA]2.0.CO;2)
47. Revell LJ. 2012 phytools: an R package for phylogenetic comparative biology (and other things). *Methods Ecol. Evol.* **3**, 217–223. (doi:10.1111/j.2041-210X.2011.00169.x)
48. Orme D, Freckleton R, Thomas G, Petzoldt T, Fritz S, Isaac N, Pearse W. 2012 caper: Comparative analyses of phylogenetics and evolution in R, v. 0.5. See <http://CRAN.R-project.org/package=caper>.
49. Vito M, Muggeo R. 2008 segmented: an R package to fit regressions model with broken-line relationships. *R News* **8**, 20–25.
50. Harmon LJ, Weir JT, Brock CD, Glor RE, Challenger W. 2008 GELER: investigating evolutionary radiations. *Bioinformatics* **24**, 129–131. (doi:10.1093/bioinformatics/btm538)
51. Pagel M. 1999 Inferring the historical patterns of biological evolution. *Nature* **401**, 877–884. (doi:10.1038/44766)
52. Blomberg SP, Garland T, Ives AR. 2003 Testing for phylogenetic signal in comparative data: behavioral traits are more labile. *Evolution* **57**, 717–745. (doi:10.1111/j.0014-3820.2003.tb00285.x)
53. Blomberg SP, Garland JT. 2002 Tempo and mode in evolution: phylogenetic inertia, adaptation and comparative methods. *J. Evol. Biol.* **15**, 899–910. (doi:10.1046/j.1420-9101.2002.00472.x)
54. Kembel SW, Cowan PD, Helmus MR, Cornwell WK, Morlon H, Ackerly DD, Blomberg SP, Webb CO. 2010 Picante: R tools for intergrading phylogenies and ecology. *Bioinformatics* **26**, 1463–1464. (doi:10.1093/bioinformatics/btq166)
55. Kotschal K, Van Staaden MJ, Huber R. 1998 Fish brains: evolution and environmental relationships. *Rev. Fish Biol. Fish.* **8**, 373–408. (doi:10.1023/A:1008839605380)
56. Gonzalez-Voyer A, Winberg S, Kolm N. 2009 Social fishes and single mothers: brain evolution in African cichlids. *Proc. R. Soc. B* **276**, 161–167. (doi:10.1098/rspb.2008.0979)
57. Gonzalez-Voyer A, Winberg S, Kolm N. 2009 Brain structure evolution in a basal vertebrate clade: evidence from phylogenetic comparative analysis of cichlid fishes. *BMC Evol. Biol.* **9**, 238. (doi:10.1186/1471-2148-9-238)
58. Huber R, van Staaden MJ, Kaufman LS, Liem KF. 1997 Microhabitat use, trophic patterns, and the evolution of brain structure in African cichlids. *Brain Behav. Evol.* **50**, 167–182. (doi:10.1159/000113330)
59. Mboko SK, Kohda M. 1995 Pale and dark dichromatism related to microhabitats in a herbivorous Tanganyikan cichlid fish, *Telmatochromis temporalis*. *J. Ethol.* **13**, 77–83. (doi:10.1007/BF02352566)
60. Kohda M, Hori M. 1993 Dichromatism in relation to the trophic biology of predatory cichlid fishes in Lake Tanganyika, East-Africa. *J. Zool.* **229**, 447–455. (doi:10.1111/j.1469-7998.1993.tb02647.x)
61. Hori M, Watanabe K. 2000 Aggressive mimicry in the intra-populational color variation of the Tanganyikan scale-eater *Perissodus microlepis* (Cichlidae). *Environ. Biol. Fishes* **59**, 111–115. (doi:10.1023/A:1007657419083)
62. Schelly R, Takahashi T, Bills R, Hori M. 2007 The first case of aggressive mimicry among lamprologines in a new species of *Lepidiolamprologus* (Perciformes: Cichlidae) from Lake Tanganyika. *Zootaxa* **1638**, 39–49.
63. Salzburger W. 2009 The interaction of sexually and naturally selected traits in the adaptive radiations of cichlid fishes. *Mol. Ecol.* **18**, 169–185. (doi:10.1111/j.1365-294X.2008.03981.x)
64. Kidd MR, Kidd CE, Kocher TD. 2006 Axes of differentiation in the bower-building cichlids of Lake Malawi. *Mol. Ecol.* **15**, 459–478. (doi:10.1111/j.1365-294X.2005.02787.x)
65. Salzburger W, Meyer A, Baric S, Verheyen E, Sturmbauer C. 2002 Phylogeny of the Lake Tanganyika cichlid species flock and its relationship to the Central and East African haplochromine cichlid fish faunas. *Syst. Biol.* **51**, 113–135. (doi:10.1080/106351502753475907)
66. Krabbenhoft TJ, Collyer ML, Quattro JM. 2009 Differing evolutionary patterns underlie convergence on elongate morphology in endemic fishes of Lake Waccamaw, North Carolina. *Biol. J. Linn. Soc.* **98**, 636–645. (doi:10.1111/j.1095-8312.2009.01305.x)
67. Boettiger C, Coop G, Ralph P. 2012 Is your phylogeny informative? Measuring the power of comparative methods. *Evolution* **66**, 2240–2251. (doi:10.1111/j.1558-5646.2011.01574.x)
68. Slater GJ, Harmon LJ, Alfaro ME. 2012 Integrating fossils with molecular phylogenies improves inference of trait evolution. *Evolution* **66**, 3931–3944. (doi:10.1111/j.1558-5646.2012.01723.x)
69. Losos JB. 2011 Seeing the forest for the trees: the limitations of phylogenies in comparative biology. *Am. Nat.* **177**, 709–727. (doi:10.1086/660020)

Chapter 9

Chapter 9.3

Testing the stages model in the adaptive radiation of
cichlid fishes in East African Lake Tanganyika

Supplementary material

Supplementary Data 1-3 to Muschick et al. (2014) can be obtained online under:
<http://rspb.royalsocietypublishing.org/content/281/1795/20140605.figures-only>

Testing the stages model in the adaptive radiation of cichlid fishes in East African Lake Tanganyika

Moritz Muschick^{1,2}, Patrik Nosil¹, Marius Roesti², Marie Theres Dittmann^{3,4},
Luke Harmon⁵ & Walter Salzburger²

Electronic Supplementary Material

Index:

Supplementary Methods

Supplementary Tables

Supplementary Figures

Supplementary Methods

Sampling

Sampling was performed between 2008 and 2012 in the southern basin of Lake Tanganyika and under the permission of the Lake Tanganyika Research Unit, Department of Fisheries, Republic of Zambia. Specimens were caught with gillnets set by snorkelling and scuba diving, by harpooning, by angling, or, in a few cases, were obtained from local fishermen. For sample preparation, we followed our standard operating procedure [described in 21]. In short, all specimens were photographed, measured and weighted; a fin-clip or a piece of muscle-tissue preserved in ethanol was taken as DNA and stable isotope sample. In total, we collected trait data for 51 LT cichlid species, which is about one quarter of the endemic species of this lake and covers 36 of the 53 genera and 10 of the 16 tribes. The choice of species was restricted to those occurring in the Southern part of the lake, where sampling took place. We used a number of sampling techniques to avoid methodical bias. Species were chosen from a larger data set solely on grounds of data completeness, i.e. data for all traits had been obtained for a given species. Efforts had been made to complete the data set also for less abundant species, in order to reduce representation bias towards more abundant species.

Trait data new to this study

The gill raker apparatus was investigated for four morphological traits: we separately counted the number of gill rakers on the dorsal and ventral bone of the first gill arch. Gill raker length was determined as the average length of rakers two, three and four on the ventral gill arch, counted from the joint with the dorsal arch bone onwards. We further measured the ventral gill arch length, as proxy for the spacing between gill rakers along the gill arch [32, 34]. These traits were measured to the nearest 0.01 mm and counted on one side within each specimen using a Leica MZ7.5 stereomicroscope. All gill raker trait assessments essentially followed previous investigations in threespine stickleback [32].

Brain tissue was removed from the neurocranium in the field and stored in either Ethanol or RNAlater (Sigma-Aldrich, Saint Louis, USA). In the laboratory, preserved fish brains were drained and dried at 60°C overnight in an incubator and subsequently weighed to the nearest milligram. A systematic bias between the two preserving liquids was apparent, with salt residues from RNAlater increasing the brain weight relative to fish body weight. We therefore adjusted the RNAlater sample measurements to fit the linear model of brain weight and body

weight for ethanol-preserved samples. To investigate gut length, we removed the entire alimentary canal ('gut') from the anus to the posterior end of the stomach from freshly caught specimens. Each gut was unwound, stretched out, and measured to the nearest millimetre.

In order to evaluate overall body colouration, we adopted and modified an existing colour-scoring scheme developed for Lake Tanganyika cichlids [41] using representative photographs. We used the 12 landmarks described in Salzburger *et al.* [41] plus information on lips (light/dark/yellow/blue), facial pattern (uniform colouration/stripes/dots), caudal fin pattern (uniform/stripes/blotches), caudal fin colouration (light/dark/yellow/orange/blue/brown/red), as well as sexual dimorphism (yes/no) and polychromatisms (yes/no). From these 101 binary-coded colouration traits we retained the 83 that showed variation across species and had each character state found in at least two species. As no information on within-species variance was available, colour was not included in analyses requiring such variation.

Data re-used from a previous study

Phylogenetic relationships were derived from the enforced molecular-clock phylogeny of Muschick *et al.* [21] by pruning it to the 51 species included in this study. This phylogeny is based on one mitochondrial marker (ND2, alignment length: 1047 bp), and two nuclear markers (*ednrb1* and *phpt*, with alignment lengths of 542 bp and 424 bp, respectively). Body shape was assessed by using Cartesian coordinates of landmarks derived from lateral, standardized photographs using TPSDIG [42]. Seventeen homologous landmarks covered the whole body and captured ecologically important shape features such as fin insertion points, body depth and length and relative size of head and trunk. Landmark coordinates were procrustes aligned in MORPHOJ [43] and, together with centroid size, analysed in R [44]. LPJ shape information was obtained similar to body shape. Twenty-eight landmark positions were recorded on scans of excised and cleaned lower pharyngeal jawbones in occlusal perspective. Twenty of these landmarks were semi-landmarks and subjected to a sliding process in TPSRELW [45] to more accurately capture the curved shapes of LPJ outlines. Of this set, 8 landmarks and 6 semi-landmarks were used for analysis. Procrustes alignment was performed in MORPHOJ and the symmetric component of shape variation and the centroid size exported for analysis in R.

Stable isotope data were used as proxies for habitat preference and trophic niche ($\delta_{13}\text{C}$ and

$\delta_{15}\text{N}$, respectively). Stable isotope ratios of carbon ($\delta_{13}\text{C}$) are related to the source of carbon in the diet and thereby reveal approximate habitat preference on a limnetic-benthic axis [46]. The heavier stable isotope of nitrogen accumulates with trophic level making $\delta_{15}\text{N}$ values can a good proxy for the trophic level of a species. Muschick *et al.* [21] used mass spectrometry on dried and pulverized white muscle tissue to reveal the isotopic composition. It is expressed in the conventional δ notation as permil (‰) deviation *versus* atmospheric N_2 and Pee Dee Belemnite.

Data transformation prior to statistical analyses

Prior to statistical analyses we log transformed all trait values, apart from landmark procrustes coordinates, gill raker counts, and colouration scores. We then regressed all trait values (excluding colouration traits) onto log transformed body weight using regression coefficients from the `phyl.resid` function of the R package PHYTOOLS version 0.2-40 [47]. Retained residuals of species means or individual trait values were used for subsequent analyses. Brain weight and gut length data were univariate, and hence treated accordingly. Data collected for the other traits were multivariate. We therefore reduced dimensionality within each trait complex using scaling, centring and principal components analysis (PCA). We calculated eigenvectors of species means of multivariate residuals using phylogenetic Principal Component Analysis (pPCA) as implemented in PHYTOOLS and retained principal components (PC) that each explained at least 10% of the total variance. Eigenvectors derived from the pPCA on species means were used to calculate PC scores of individuals. Residuals of univariate traits were scaled and centred only. As the colouration matrix consisted of binary coded data, colouration scores were reduced in dimensionality using detrended correspondence analysis. We kept the first three detrended correspondence axes (DFA) for further analyses.

Ecological specialization and overlap between species

We used plots of linear discriminants (LD) to illustrate each species' position in morphospace for the multivariate data (body and lower pharyngeal jaw shape, gill rakers), and boxplots to illustrate the univariate data (gut length, brain weight). In the first case, we plotted LD1 against LD2, representing each species as polygons defined by the most extreme individuals,

to illustrate the subdivision of the morphospace among species and to indicate niche overlap between species. The boxplots show the log transformed, regressed, scaled, and centred trait values for each species, sorted by median.

Next, we calculated the between-species distances for each trait. Traits for which within-species variance information was available were used to calculate the Mahalanobis distance from the within-species covariance matrix. Since Mahalanobis distances are scaled by the within-group variance, we used them as a generalized measure of trait divergence, which can be compared among traits. Conceptually, this approach is similar to measuring evolutionary rate in *haldanes* by using the pooled standard variation to standardize character change over time [S1].

Since no such data were available for colour, we used the absolute (Manhattan) distance in the ordinated colour space derived from the presence/absence colour trait matrix. Distance matrices from different traits were then tested for correlations between each other using three-way partial Mantel tests, correcting for phylogenetic distance. The phylogenetic distance matrix was calculated using the cophenetic function in R on the molecular phylogeny. Significance of the Mantel statistic was tested with 9999 permutations.

Supplementary Reference:
Gingerich PD. 1993 Quantification and comparison of evolutionary rates. *Am. J. Sci.* 293, 453–478.
(doi:10.2475/ajs.293.A.453)

Table S1 Number of specimens measured per species per trait.

species	abbreviation	body shape	LPJ shape	gut length	gill raker	brain weight	stable isotopes
<i>Altalamprologus calvus</i>	altcal	6	5	2	1	1	0
<i>Altalamprologus compressiceps</i>	altcom	23	10	12	11	10	10
<i>Altalamprologus fasciatus</i>	altfas	13	11	8	10	8	11
<i>Asprotilapia leptura</i>	asplep	11	7	7	8	1	10
<i>Astatotilapia burtoni</i>	astbur	9	11	10	10	3	7
<i>Aulonocranus dewindtii</i>	auldew	36	19	9	11	5	16
<i>Benthochromis tricoti</i>	bentri	6	6	5	3	3	7
<i>Boulengerochromis microlepis</i>	boumic	17	6	2	2	2	10
<i>Callochromis macrops</i>	calmac	16	8	9	10	1	10
<i>Chalinochromis brichardi</i>	chabri	7	9	4	4	3	9
<i>Cyphotilapia gibberosa</i>	cphgib	15	12	8	8	5	13
<i>Ctenochromis horei</i>	ctehor	31	13	11	10	9	17
<i>Cyathopharynx furcifer</i>	cyafur	45	34	10	10	5	11
<i>Cyprichromis leptosoma</i>	cyplep	16	12	10	10	6	11
<i>Enantiopus melanogenys</i>	enamel	20	6	7	7	7	10
<i>Eretmodus cyanostictus</i>	erecya	15	15	4	8	5	14
<i>Gnathochromis pfefferi</i>	gnapfe	13	8	7	8	8	11
<i>Grammatotria lemairii</i>	gralem	20	15	4	4	5	13
<i>Haplotaxodon microlepis</i>	hapmic	15	9	4	4	8	10
<i>Haplotaxodon trifasciatus</i>	haptri	4	5	11	11	5	4
<i>Interochromis loocki</i>	intloo	10	12	11	11	9	10
<i>Julidochromis ornatus</i>	julorn	11	10	8	8	8	10
<i>Lamprologus callipterus</i>	lamcal	21	13	10	12	8	11
<i>Lamprologus lemairii</i>	lamlem	13	12	5	5	6	13
<i>Lepidiolamprologus attenuatus</i>	lepatt	26	18	10	10	8	13
<i>Lepidiolamprologus elongatus</i>	lepelo	34	21	10	10	10	19
<i>Lepidiolamprologus profundicola</i>	leppro	8	10	5	5	4	8
<i>Limnotilapia dardenni</i>	limdar	29	22	7	7	7	12
<i>Lobochilotes labiatus</i>	loblab	31	14	14	15	10	13
<i>Neolamprologus furcifer</i>	neofur	13	8	1	1	1	9
<i>Neolamprologus modestus</i>	neomod	25	17	12	12	11	13
<i>Neolamprologus pulcher</i>	neopul	21	10	9	11	8	10
<i>Neolamprologus savoryi</i>	neosav	22	10	11	11	5	9
<i>Neolamprologus sexfasciatus</i>	neosex	13	11	8	8	6	10
<i>Neolamprologus tetracanthus</i>	neotet	20	23	6	6	7	11
<i>Ophthalmotilapia ventralis</i>	ophven	17	14	12	11	11	20
<i>Paracyprichromis brieni</i>	pcybri	5	7	4	5	5	10
<i>Perissodus microlepis</i>	permic	16	30	10	10	10	11
<i>Petrochromis ephippium</i>	peteph	12	10	8	8	10	10
<i>Petrochromis famula</i>	petfam	10	10	7	7	4	12
<i>Petrochromis macrognathus</i>	petmac	18	12	12	11	10	11
<i>Petrochromis polyodon</i>	petpol	10	14	7	7	3	10
<i>Plecodus straeleni</i>	plestr	11	9	5	10	9	10
<i>Pseudosimochromis curvifrons</i>	pscscr	13	8	9	10	11	10
<i>Simochromis diagramma</i>	simdia	27	13	10	10	10	17
<i>Telmatachromis temporalis</i>	teltem	11	16	10	10	5	9
<i>Trapheus moori</i>	tromoo	28	16	10	10	10	21
<i>Tylochromis polylepis</i>	tylpol	11	9	3	3	4	10
<i>Variabilichromis moori</i>	varmoo	23	21	10	10	10	18
<i>Xenotilapia flavipinnis</i>	xenfla	7	9	10	10	8	9
<i>Xenotilapia spiloptera</i>	xenspi	32	19	3	5	6	14
sum	51	886	649	401	419	334	577
mean	-	17.4	12.7	7.9	8.2	6.5	11.3
median	-	15	11	8	10	7	10

Table S2. Statistical analyses of species' separation in morphospace. MANOVA and ANOVA results for multivariate traits and univariate traits, respectively.

trait	approx. F	Df1	Df2	Pr (>F)	Wilks' λ
MANOVA					
body shape	65.35	50	835	0	0.0084
lpj shape	53.95	50	598	0	0.0059
gill raker	59.51	50	368	0	0.0121
ANOVA					
gut length	38.29	50	350	0	-
brain weight	21.58	50	283	0	-

Table S3. Macroevolutionary models fitted to a molecular phylogeny and trait data of Tanganyikan cichlid fishes. Parameter estimates and quality of model fit for macroevolutionary models, and estimation and significance of Blomberg's K. Sample-size corrected Akaike Information Criterion (AICc) indicates model fit, with a smaller score being a better fit. The best fitting model for each trait among the evolutionary process and rate variation models is indicated by italicized AICc score.

	body shape			lpoj shape			gill raker		brain	gut	colour		
	pPC 1	pPC 2	pPC 3	pPC 1	pPC 2	pPC 3	pPC 1	pPC 2	residuals	residuals	DCA 1	DCA 2	DCA 3
Evolutionary process													
Brownian motion (σ^2)	252.44	103.95	82.21	175.05	136.19	86.05	57.57	11.34	14.04	17.28	13.22	5.64	1.02
AIC _c	269.91	231.76	222.02	248.05	242.35	210.69	180.93	113.45	123.93	116.65	169.07	156.07	156.66
Ornstein-Uhlenbeck (α)	16.39	0	35.5	15.98	32.22	21.25	0	0	0	7.01	0	13.18	50.96
AIC _c	270.85	230.76	218.85	248.93	240.20	208.79	181.88	115.82	126.3	118.43	171.44	158.38	157.58
white noise (σ^2)	11.76	5.53	3.97	10.12	6.69	4.42	2.64	0.67	0.98	0.98	1.64	1.16	1.11
AIC _c	276.94	238.48	221.52	269.31	248.21	227.1	200.71	131.08	150.24	150.24	176.5	158.83	156.42
Rate variation													
Pagel's δ	6.35	3.83	5.20	3.02	4.89	3.36	2.26	1.42	1.31	1.86	1.45	2.12	20.06
AIC _c	275.92	230.89	218.55	248.52	240.5	208.65	181.93	120.89	138.09	118.12	171.27	158.17	158.04
Early Burst (a)	32.82	50.55	70.99	31.97	64.44	42.5	19.44	1	1	14.02	4194.8	26.37	101.91
AIC _c	270.85	231.61	218.85	248.93	240.20	208.79	182.47	115.96	126.54	118.43	178.91	158.38	157.58
Phylogenetic signal													
Pagel's λ	0.88	0.88	0.44	0.94	0.83	0.95	0.99	0.91	0.94	1.00	0.59	0.38	0.00
AIC _c	268.45	233.47	223.07	250.42	242.64	213.06	183.29	115.81	126.3	119.02	171.44	158.44	158.79
Blomberg's K	0.41	0.34	0.44	0.67	0.48	0.63	0.82	0.76	1.02	0.96	0.21	0.33	0.32
P value	0.011	0.06	0.003	0.001	0.002	0.001	0.001	0.001	0.001	0.001	0.576	0.077	0.257

Figure S1. Gill raker structures on a ventral gill arch of *Simochromis diagramma* (tribe Tropheini).



Figure S2. Phenotype-environment correlations. Species means for PC 1 (or DCA 1, in the case of colour) of traits was plotted against species means of $\delta_{13}\text{C}$, $\delta_{15}\text{N}$, and the first principal component of both. $\delta_{13}\text{C}$ is considered to be an indicator of macrohabitat choice, while $\delta_{15}\text{N}$ is a proxy for trophic level.

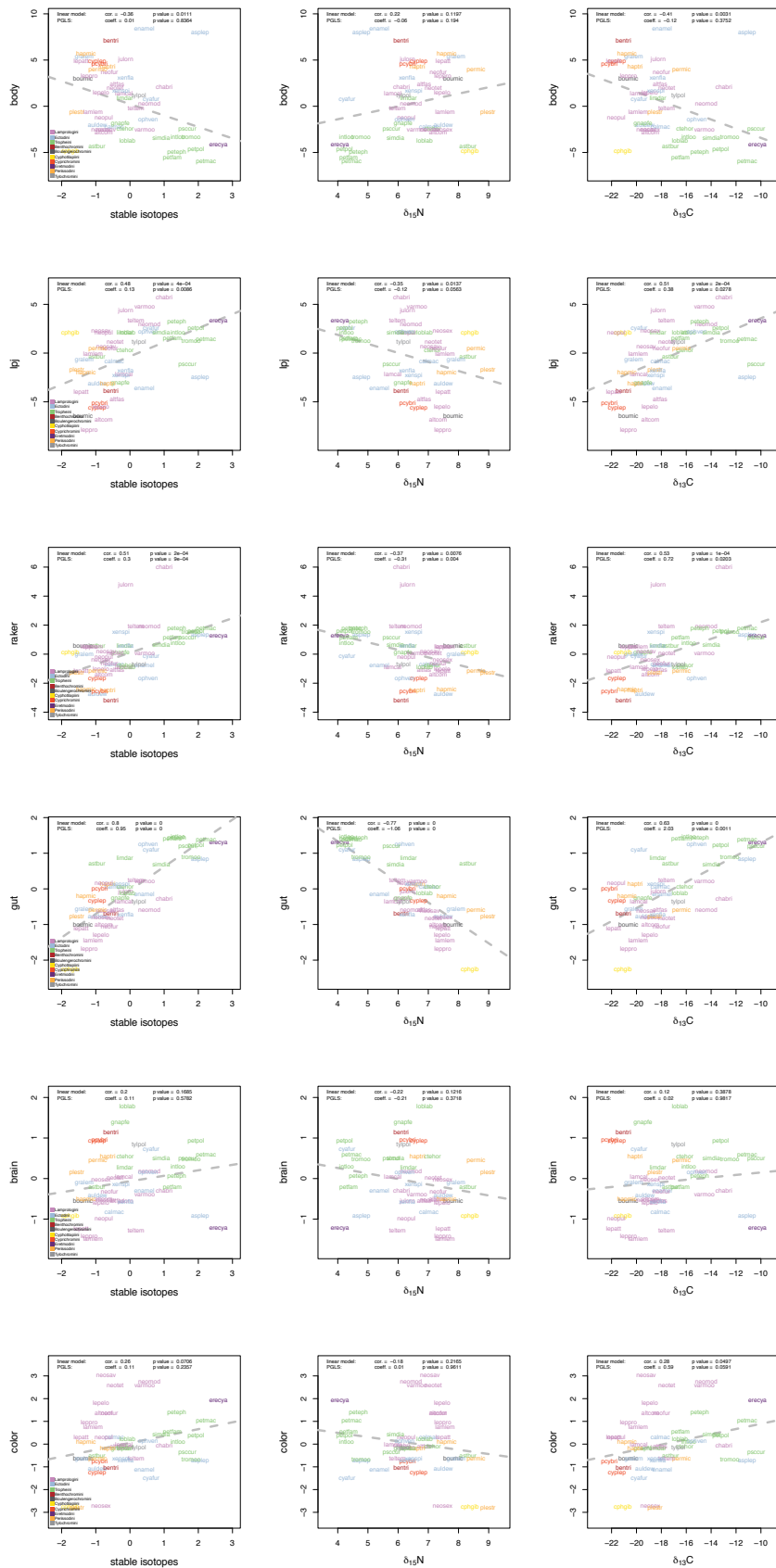
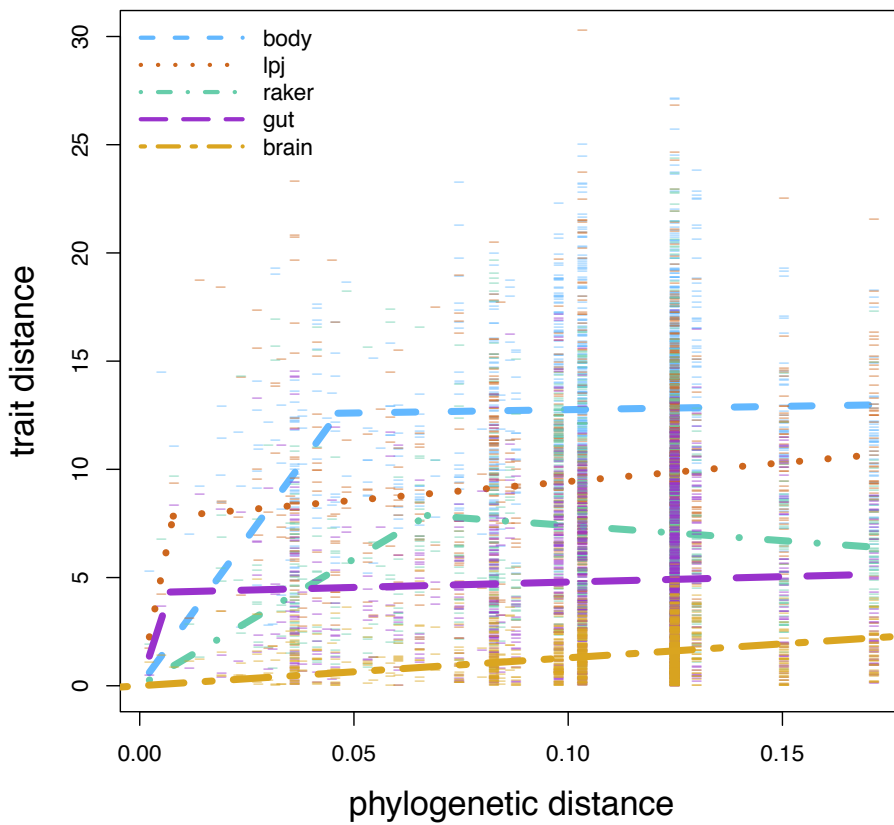


Figure S3. Breakpoint regressions of trait distances on phylogenetic distance. Between-species Mahalanobis distances for multivariate traits, and Manhattan distances scaled by within-species variance plotted against phylogenetic distance derived from the ultrametric phylogeny of [21]. Breakpoint regression models were fitted for each trait, partitioning the correlation of distances into two phases of trait-divergence build-up. A linear model without breakpoint was the best fit for brain weight distances. The lack of correlation in the second phase indicates little change in average between-species trait distances over evolutionary time for all traits. Attained average trait distances can be compared across traits as they are standardized by within-species variance.



Thesis Conclusion

This thesis leads to general insights into the genomics of adaptation that exceed the discussions in the different main chapters. In what follows, I will address these insights as well as their implications (I here focus mainly on the Main Chapters).

Genome Divergence is Heterogeneous

A general finding of this thesis is that divergence between populations in the face of gene flow is highly heterogeneous along the genome (Chapters 1, 3, 4, 6, 7). This implies that the strength of genetic isolation between populations also differs along the genome. Genomic heterogeneity in divergence can be found at various scales such as along entire chromosomes (Chapters 1, 3, 4), or within only a few kilobases (Chapters 1, 4, 6). Besides selection, variation in recombination within the genome is important to explain some of this heterogeneity (Chapter 3; Butlin 2005). Importantly, the detected physical extent and intensity of a divergence region does not necessarily reflect the strength with which divergent selection has acted on that region. This is because large adaptive structural rearrangements (e.g., inversions, Chapters 4 & 6) usually appear prominently in genome scans and can already be detected with a limited number of markers. In contrast, an equally strong (or stronger) selected single point mutation is likely to be missed when we screen a genome. The main reason for this ‘detection problem’ is that the extent and intensity of a selective sweep around a selected locus strongly depends on its location within the genome (e.g., whether it is close to other selected loci, or located in a high/low recombination region) (Chapters 1, 3, 4). A bias in the occurrence of cross-overs along chromosomes together with selection can result in an uneven ‘neutral’ divergence-baseline along chromosomes (Chapters 1 & 3). This heterogeneity in baseline divergence has to be accounted for when we are interested in the inference of actual *targets* and the *strength* of selection along the genome (Chapters 1 & 3). The intensity and extent of selective sweeps also strongly depend on the type of genetic variation available to selection (i.e., ancestral standing vs. young *de novo* variation) and on the time selection has acted. As an additional complication, genomic divergence hot spots may not necessarily indicate a region under divergent selection (Chapter 4). Overall, I conclude that patterns of genomic divergence can only be understood within a rigorous ecological framework and with extensive knowledge about the demographic and selective history of populations.

Adaptation is Complex

My thesis work adds evidence from different methodological angles (Chapters 1, 3-6, 8) to the emerging perspective that adaptation is a complex multi-locus process (e.g., Burke et al. 2010; Lawniczak et al.

2010; Fournier-Level et al. 2011; Soria-Carrasco et al. 2014): A first reason is that adaptation as a whole is likely to require shifts at many phenotypic traits, including behavior, morphology, physiology, and life history (Chapters 8 & 9). At a second level, even single ecologically relevant traits are commonly under the control of many genetic loci with small phenotypic effects (Chapter 5). This concept of adaptation as a complex process involving many traits and loci with small phenotypic effects has important implications:

(i) Our theoretical understanding of the localized and genome-wide signatures of selection proves incomplete. This is because our expectations of how selection should become manifest in the genome is still much influenced by theoretical insights from (historic) single-locus models. Clearly, we should generate new theory that explicitly incorporates our most recent understanding from empirical evolutionary genomics research (see e.g., Chapter 4; Flaxman et al. 2014).

(ii) The characterization of some genes with a large phenotypic effect in population divergence has put a strong research focus on finding ‘speciation genes’ (Orr et al. 2004; Nosil & Schluter 2011). Intriguingly, this trend seems to bias our perception of the complexity of adaptation, and to impair our scientific methodology and terminology. For example, studies often classify genetic polymorphisms (markers) either as ‘outliers’ or ‘non-outliers’. However, in the light of our current understanding of the genomics of adaptation, such classification is misleading as it implies the possibility to clearly distinguish between neutrally and non-neutrally evolving loci (Butlin 2010). For instance, there is no indication for a biologically justified threshold in the distribution of F_{ST} values between populations to distinguish between such distinct marker classes. Also, if hundreds of loci underlie the divergence of populations, one can question if the category ‘speciation gene’ is valid and useful.

(iii) Studies have recently tried to estimate the fitness effect of single genes or traits (e.g., Barrett et al. 2008; Terekhanova et al. 2014). Surprisingly, these studies often find high selection coefficients (up to 0.5) for single loci or traits. Nevertheless, these estimates provide us with only little information about the selective situation in the wild (in fact, such estimates are likely biased). The reason is that a fitness effect of a single locus or trait is never independent of the rest of an organism in the wild – that is, it is never independent from all other loci and traits relevant for adaptation.

Genomic Parallelism in Adaptation

To what extent is parallel phenotypic evolution mirrored by parallel genetic changes within a genome? To this point, there is no clear answer to this fundamental question. Stickleback repeatedly use pre-existing genetic variation to adapt to similar environments (Chapters 4 & 6). The quantification of the degree of parallelism at the genome level, however, remains challenging and imposes high requirements on the quality and amount of data (see below). More fundamentally, our perception of parallelism is probably biased: those loci identified with the highest certainty as loci under selection are often necessarily those displaying parallelism, because adaptation genomics studies often use parallelism as a confidence criterion (e.g., Chapter 4; Jones et al. 2012). Furthermore, the loci recycled by selection with high fidelity might be of relatively large effect and hence easiest to detect. It is also unclear to which degree our findings in stickleback allow for general conclusions, as most other non-genetic model organisms do not (yet) provide comparable data sets.

Another crucial point concerns our thinking about ‘parallelism’ as such. Can evolution still be said to be ‘parallel’ when it is not exactly the same mutation that affects a phenotype, but different mutations in the same gene (e.g., Linnen et al. 2013) or the same genetic pathway (e.g., Manceau et al. 2010)? Chapter 6 of this thesis also demonstrates a striking case of repeated selection of globally shared variants in different ecological contexts. This provides evidence for genetic parallelism in the absence of ecological parallelism (notably, the ecologically different environments might well have shared selection pressures). Another important question to be considered concerning parallelism is whether we are looking at causal variants or at neutral markers, as the latter might often not be parallel while the former are (Chapter 4).

Complementary Approaches and a Reference Genome

My research emphasizes the need to integrate conceptually and analytically complementary approaches in adaptation genomics research. QTL and association mapping (Chapters 5 & 6), for example, start with information at the phenotype-level and work down to the genotype. This top-down approach provides a strong test for causality between a phenotypic trait and certain genomic regions (Chapter 8). Nevertheless, QTL and association mapping alone do not tell us much about the degree to which a phenotype and its underlying genetics are relevant to adaptation. In contrast to top-down investigations, bottom-up approaches can identify putatively adaptive loci by searching for molecular signatures of selection in the absence of phenotype information. The archetypal bottom-up approach is divergence mapping in natural populations (Chapters 1 & 6). Bottom-up approaches generally provide a better picture of selection at the genomic level than top-down approaches, since they are not biased by focusing on a few measurable traits only. Once selected genomic regions are identified, however, inferring their function in the absence of phenotypic information is very challenging. Nevertheless, the combination of mapping approaches with bottom-up methods (and phenotypic investigations in wild populations) allows us to link genetics, phenotypes, and selection (fitness) at least for some loci. In this thesis, I take on this challenge (Chapters 5 & 6), where either QTL mapping or association mapping is combined with genomic data from natural populations. Furthermore, we should integrate complementary analytical tools in adaptation genomics research. The reason is that there is no ideal single measure or statistics to detect footprints of selection at the molecular level (Thesis Introduction; Oleksyk et al. 2010). For instance, LD-based tests are most useful for the identification of recent, incomplete selective sweeps (Chapter 6), whereas divergence tests are powerful to detect genomic regions where different allelic variants are (nearly) fixed between populations (Chapters 1, 2, 4, 6, 7). In summary, we should combine different approaches (e.g., QTL with divergence mapping, Chapter 5) as well as tests (e.g., LD with divergence tests, Chapter 6) to address quantitative and qualitative aspects of selection within the genome. This is challenging because different approaches and tests rely on different kinds of data. The quantitative inference of selection at the genome-level in wild populations requires particularly high bp-resolution, sequence coverage, and a decent number of study individuals and replicate populations.

This thesis work also clearly highlights the importance of a reliably assembled reference genome. For example, inferences about recombination (Chapter 3) or linkage (Chapters 4 & 6) demand knowledge about the physical position of markers along a genome. Otherwise the powerful integration of single point-estimates of divergence through sliding window approaches is impossible (as applied in Chapters 1-6). In brief, most questions of greatest interest in adaptation genomics (see Thesis Introduction) cannot be addressed in the absence of a well-assembled reference genome

Speciation from the Genomics Perspective

Speciation can be considered a continuous process covering a range from 'no' to 'complete' reproductive isolation (Nosil 2012). Evolutionary genomics casts a new light on this process by quantifying the extent of genomic divergence to infer the strength of isolating barriers between populations and species. The heterogeneity in genome divergence suggests that gene flow between populations happens more easily in some genomic regions than in others. This indicates that the extent of reproductive isolation between populations and species can vary strongly along the genome, adding further complexity to the long-standing discussion of how to define a 'species' and a 'population' (Mallet 1995; De Queiroz 2007).

The moment we study different populations, we can only speculate whether or not they will move further along the speciation continuum in the evolutionary future. The concept of speciation as a continuum or even as a process with distinct 'stages' (Feder et al. 2012) is delicate, since it implies a certain directionality in the course of evolution. Importantly, however, adaptive population divergence with ongoing gene flow

may itself constitute an evolutionary stable state. Even though divergence might progress directionally at the phenotype level, it is well conceivable that small and fluctuating selection coefficients of the many adaptive loci might never lead to fixed differences between some species (Buerke et al. 2010).

'Understanding' the Genomics of Adaptation and Future Perspectives

Many of the sequencing, molecular and bioinformatics techniques and approaches used in genomics research have just been developed and are thus still in their infancy. Consequently, most current evolutionary genomics studies are yet explorative (but see Chapters 3 & 4), and many of our conceptual and methodological ideas and challenges are stimulated through (unexpected) empirical findings (Chapters 1-3). Nevertheless, the progress in our understanding of the genetics and genomics of adaptation, and evolution in general, has been overwhelming in the past decade.

A major challenge for the field of evolutionary genomics is how to deal with the genomic complexity of adaptive evolution and the many associated methodological and conceptual challenges. It thus seems necessary to consider (i) how much molecular detail we need to illuminate for an adequate understanding of evolution, and (ii) our inherent limits in understanding and interpreting genomic patterns. Because evolutionary biology is in essence a historical science, we have to concede that the identification of all causative processes underlying present-day (genomic) patterns is likely to be impossible. At the same time, general principles can only emerge by integrating many different and replicate case studies from various organisms. At its best, this thesis makes such a contribution and adds to our general understanding of the genomics of adaptation and evolution in general.

References

- Barrett SCH, Rogers SM, Schluter D (2008) Natural selection on a major armor gene in threespine stickleback. *Science* 322, 255-257.
- Burke MK, Dunham JP, Shahrestani P, et al. (2010) Genome-wide analysis of a long-term evolution experiment with *Drosophila*. *Nature* 467, 587-590.
- Butlin RK (2005) Recombination and speciation. *Molecular Ecology* 14, 2621-2635.
- Butlin RK (2010) Population genomics and speciation. *Genetica* 138, 409-418.
- De Queiroz K (2007) Species concepts and species delimitation. *Systematic Biology* 56, 879-886.
- Feder JL, Egan SP, Nosil P (2012) The genomics of speciation-with-gene-flow. *Trends in Genetics* 28, 342-350.
- Flaxman SM, Wacholder AC, Feder JL, Nosil P (2014) Theoretical models of the influence of genomic architecture on the dynamics of speciation. *Molecular Ecology* 23, 4074-4088.
- Fournier-Level A, Korte A, Cooper MD, et al. A map of local adaptation in *Arabidopsis thaliana*. *Science* 334, 86-89.
- Jones FC, Grabherr MG, Chan YF, et al. (2012) The genomic basis of adaptive evolution in threespine sticklebacks. *Nature* 484, 55-61.
- Lawniczak MKN, Emrich SJ, Holloway AK, et al. (2010) Widespread divergence between incipient *Anopheles gambiae* species revealed by whole genome sequences. *Science* 330, 512-514.
- Linnen CR, Poh Y-P, Peterson BK, et al. (2013) Adaptive Evolution of Multiple Traits Through Multiple Mutations at a Single Gene. *Science* 339, 1312-1316.
- Mallet J (1995) A species definition for the modern synthesis. *Trends in Ecology & Evolution* 10, 294-299.
- Manceau M, Domingues VS, Linnen CR, Rosenblum EB, Hoekstra HE (2010) Convergence in pigmentation at multiple levels: mutations, genes and function. *Philosophical Transactions of the Royal Society of London Series B* 365, 2439-2450.
- Nosil P (2012) *Ecological Speciation*, 1 edn. Oxford University Press, New York.
- Nosil P, Schluter D (2011) The genes underlying the process of speciation. *Trends in Ecology and Evolution* 26, 160-167.
- Oleksyk TK, Smith MW, O'Brien SJ (2010) Genome-wide scans for footprints of natural selection. *Philosophical Transactions of the Royal Society of London Series B* 365, 185-205.
- Orr HA, Masly JP, Presgraves DC (2004) Speciation genes. *Current Opinion in Genetics & Development* 14, 675-679.
- Soria-Carrasco V, Gompert Z, Comeault AA, et al. (2014) Stick Insect Genomes Reveal Natural Selection's Role in Parallel Speciation. *Science* 344, 738-742.
- Terekhanova NV, Logacheva MD, Penin AA, et al. (2014) Fast evolution from precast bricks: genomics of young freshwater populations of threespine stickleback *Gasterosteus aculeatus*. *PLoS Genetics* 10, e1004696.

Acknowledgements

First of all, I would like to thank Walter Salzburger. He gave me an incredible scientific freedom during my time as a PhD student. This allowed me to pursue my own ideas, to be creative, and also to keep on working with stickleback. Walter opened as many doors for me as he possibly could and supported me throughout my studies. I am thankful to him for giving me the chance to take part in various different projects, which provided me with valuable insights into working with other organisms and brought me in contact with many interesting people. The field trips to Zambia (and prior to my PhD studies to Central America) I count to my most memorable experiences ‘above and under water’! I think that all of this had not been possible if Walter didn’t believe in what I was doing for the past years; for all of this, Walter, I am truly thankful.

I am most thankful to Daniel Berner. Daniel has been an incredible mentor during the past years and he has been my most inspiring scientific source. He has led by example how to be a happy, conscientious, modest, and truly outstanding scientist. Daniel taught me how to ask intelligent biological questions and how to integrate multidisciplinary approaches to answering them. I have greatly benefited from his insightful feedback and support at every stage of my work. The scale and scope of this thesis would not have been possibly without him. I count myself as incredibly lucky to know Daniel, as a scientist and as a person. Thank you Daniel!

I would like to thank all my collaborators. It would go too far to mention everyone, but some deserve to be pointed out: I am thankful to Andrew Hendry for his constant feedback on my work (and for the fun times we had together in Montreal). My manuscripts have greatly profited from his mastery of language. I am honored that Sergey Gavrilets gave me the opportunity to collaborate with him. Dario Moser has contributed with many working hours to various projects of this thesis and our field-trips have always been a pleasure! Michael Matschiner has been a continuous source for solutions to computer-related problems. Also, I thank Anya Theis for her generous help with compiling my final thesis as a single document.

I had great office mates throughout my PhD! They skillfully managed to be (mostly) quiet and great fun at the same time. Thank you Eveline and Sara (the ladies sitting right next to me), and thank you as well Dario, Marco and Nico (slightly further away, but still breathing the same office-air).

Along the way, I have met special people, two of whom deserve to be mentioned here. Thank you Katie for hosting me at your place and for coming to Basel during your sabbatical. – Jacquelin, you are simply incredible.

The time in ‘Walter’s group’ will remain unforgettable to me! A list with names and reasons for this would very likely be incomplete and not justify the things I take with me from this time. I am happy to share many great memories with so many of you!

Finally, special thanks go to Petra and my family, my most important people ‘behind the scenes’. Thank you for always being there for me.

