The Genetic Basis of Adaptation and Speciation in Benthic and Limnetic Threespine Stickleback (*Gasterosteus aculeatus*)

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von
Muhua Wang
aus Guizhou, China

Tübingen

2018

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation: 06.07.2018
Dekan: Prof. Dr. Wolfgang Rosenstiel
1. Berichterstatter: Dr. Felicity Jones
2. Berichterstatter: Prof. Dr. Nico Michiels

CONTRIBUTIONS

In the study that I presented here in the thesis, I designed the experiments, performed all the data analyses, generated all the figures, and wrote the thesis.

My advisor, Dr. Felicity Jones, conceived the original idea of the project, collected and whole genome sequenced several stickleback individuals used in this study, contributed to the experimental design, guided my analyses throughout the project, and provided suggestions for the thesis writing.

My colleague, Ms. Li Ying Tan, performed the functional dissection experiments described in Chapter 5.

My colleague, Ms. Vrinda Venu, constructed whole genome re-sequencing libraries for several stickleback individuals.

My colleague, Dr. Jukka-Pekka Verta, collected and constructed whole genome sequencing libraries for several stickleback individuals. Dr. Verta also collected and constructed all the RNA sequencing libraries.

The members of Schluter Lab at the University of British Columbia collected the samples of benthic and limnetic sticklebacks.

Our collaborators across the Northern Hemisphere collected the samples of marine and freshwater sticklebacks.

The sequencing team at the genome center of the Max Planck Institute for Developmental Biology performed whole genome sequencing and RNA sequencing for all the samples.

ACKNOWLEDGEMENT

First and foremost I want to thank my advisor Felicity Jones for giving me the opportunity to study in her lab. Without her constant help and wise suggestions, it is impossible for me to finish my thesis research. She has showed me the beauty of evolutionary biology and taught me how good research in adaptive evolution was done. I appreciate all her contributions of time and ideas to make my Ph.D. study productive and unforgettable.

I want to express my sincere thanks to my secondary advisor, Prof. Dr. Nico Michiels, for his valuable suggestions of my research and good discussions about fish biology.

I want to thank to Dr. Frank Chan for accepting to serve on my Ph.D. scientific advisory committee and his valuable suggestions throughout my Ph.D. study. I would like to thank Dr. Detlef Weigel, another member of my Ph.D. scientific advisory committee, for his advice and input.

I would like to thank Prof. Dr. Oliver Betz and Prof. Dr. Ralf Sommer for accepting to be two other members of my Ph.D. examiner committee.

I want to use this opportunity to thank Dr. Dolph Schluter, my collaborator from the University of British Columbia, for providing the fish samples used in this study and valuable suggestions throughout my Ph.D. study. Dr. David Kingsley from Stanford University whose contribution to genomic sequencing data sets is greatly appreciated. And a large number of people in stickleback community across Northern Hemisphere for collecting marine and freshwater stickleback samples used in this study.

I want to thank the members of Jones lab: Li Ying Tan, for performing all the functional dissection experiments in this study and her suggestions for my thesis writing; Vrinda Venu, for performing whole genome re-sequencing for several individuals used in this study; Jukka-Pekka Verta, for performing RNA sequencing in my study and his advice on allele specific expression analysis; Andreea Dréau, for the her suggestions of computer programing; Melanie Kirch, for translating my abstract to German; Stanley Neufeld, for his suggestions and buddy talk throughout my Ph.D, Elena Avdievich, for managing the lab so nicely; Enni Harjunmaa, for her valuable suggestions of functional dissection experiments; Sebastian Kick, for taking care of our lovely fishes.

And Chan lab:

Michelle Yancoskie and Layla Hiramatsu, for their suggestions of my thesis writing; Marek Kucka, Stefano Lazzarano, Ludmila Gaspar, João Castro, Alessandra Aleotii for the helpful discussions.

Last but not least, I want to express my special thanks to my mom, dad, and my angel, little “RanRan”, for letting me know what joy and happiness is.

And especially to Pengli, for showing me what love is, telling me the beauty of everyday, and supporting me whenever it is needed.

ABBREVIATIONS

Stickleback Populations

PAXB	Paxton Lake benthic sticklebacks
PAXL	Paxton Lake limnetic sticklebacks
PRIB	Priest Lake benthic sticklebacks
PRIL	Priest Lake limnetic sticklebacks
QRYB	Little Quarry Lake benthic sticklebacks
QRYL	Little Quarry Lake limnetic sticklebacks
ENSB	Enos Lake benthic sticklebacks
ENSL	Enos Lake limnetic sticklebacks
LITC_DWN	Marine sticklebacks from Little Campbell River, Canada
LITC_UP	Freshwater sticklebacks from Little Campbell River, Canada
BNMA	Marine sticklebacks from Bonsall Creek, Canada
BNST	Freshwater sticklebacks from Bonsall Creek, Canada
BIGR_DWN	Marine sticklebacks from Big River, California, USA
BIGR_UP	Freshwater sticklebacks from Big River, California, USA
MIDF_DWN	Marine sticklebacks from Midfjardara River, Iceland
MIDF_UP	Freshwater sticklebacks from Midfjardara River, Iceland
TYNE_DWN	Marine sticklebacks from River Tyne, Scotland
TYNE_UP	Freshwater sticklebacks from River Tyne, Scotland

Genes

G6PD	*Glucose-6-phosphate dehydrogenase*
LCT	*Lactase*
Mc1r	*Melanocortin-1 receptor*
Eda	*Ectodysplasin*
Pitx1	*Pituitary homeobox transcription factor 1*
Kitlg	*Kit ligand*
GDF6	*Growth/Differentiation Factor 6*
Bmp6	*Bone morphogenetic protein 6*
SCUBE1	*Signal peptide-CUB domain-EGF-related-1*
COL24A1	*Collagen type XXIV alpha1*
AR	*Androgen receptor*
MSNA	*Moesin a*
PDE4BA	*Phosphodiesterase 4B, cAMP-specific a*
HMCN1	*Hemicentin 1*
USP25	*Ubiquitin Specific Peptidase 25*
MED13A	*Mediator Complex Subunit 13a*
WNT5A	*Wingless-type MMTV integration site family, member 5a*
NAV3	*Neuron navigator 3*
EBNA1BP2	*ENBA1 binding protein 2*

OPA1	*Optic atropy 1*
DNMT3BB.2	*DNA (cytosine-5-)-methyltransferase 3 beta, duplicate b.2*
GDPD5A	*Glycerophosphodiesterase domain containing 5a*
B4GALNT1A	*Beta-1,4-N-acetyl-galactosaminyl transferase 1a*
TNNT2A	*Troponin T type 2a*
LARP7	*La ribonucleoprotein domain family, member 7*
AUTS2A	*Autism susceptibility candidate 2a*
ACOX3	*Acyl-CoA oxidase 3, pristanoyl*
SOCS3	*Suppressor of cytokine signaling 3a*
STAT3	*Signal transducer and activator of transcription 3*

Others

SNP	Single nucleotide polymorphism
N_e	Effective population size
s	Selection coefficient
AFS	Allele frequency spectrum
CLR	Composite likelihood ratio
EHH	Extended haplotype homozygosity
F_{ST}	Fixation index
LD	Linkage disequilibrium
iHS	Integrated haplotype score
SNP	Single nucleotide polymorphism
XP-EHH	Cross-population extended haplotype homozygosity
XP-CLR	Cross-population composite likelihood ratio test
BDMI	Bateson-Dobzhansky-Mueller incompatibility
QTL	Quantitative trait locus
PCA	Principal component analysis
PC1	First principal component
PC2	Second principal component
π	Nucleotide diversity
CSS	Cluster separation score
FDR	False discovery rate
GO	Gene ontology
PBS	Population branch statistic
ML	Maximum likelihood
MAF	Minor allele frequency
C.I.	Confidence Intervals
ASE	Allele-specific expression
RNA-Seq	RNA sequencing
EST	Expressed sequence tags
GFP	Green fluorescent protein

Abstract

Sympatric benthic and limnetic stickleback fishes have been independently evolved in five lakes in British Columbia, Canada. The benthic and limnetic stickleback ecotypes showed parallel divergence in morphology due to adaptation to contrasting environmental niches. The parallel evolution of benthic and limnetic stickleback ecotypes in all five lakes makes them an excellent model to study the roles of natural selection in speciation and adaptation. Although the ecology of benthic and limnetic stickleback speciation and adaptation has been intensively studied, the genetic basis of their speciation and adaptation is still lacking.

I used whole genome re-sequencing to study the speciation and adaptation of benthic and limnetic sticklebacks from four lakes in British Columbia, Canada (Paxton Lake, Priest Lake, Little Quarry Lake, Enos Lake). Benthic and limnetic sticklebacks from all four lakes show parallel genetic divergence. Benthic and limnetic stickleback ecotypes have been subject to strong divergent natural selection, in which derived alleles and ancestral alleles are selectively favored in benthic and limnetic ecotypes respectively. There are substantially more genomic regions that were selected in benthic ecotypes than limnetic ecotypes. I identified the genomic regions which contribute to the adaptation of benthic and limnetic ecotypes with unprecedented resolution by combining several statistical approaches. This allows me to identify and characterize genes controlling important adaptive phenotypic traits and biological pathways that are important for adaptation of benthic and limnetic ecotypes. Using high-density genetic markers generated from whole genome re-sequencing, I investigated the ancestry of benthic and limnetic ecotypes and inferred the demographic model of Paxton Lake benthic and limnetic sticklebacks. Paxton Lake benthic and limnetic sticklebacks were evolved from allopatric speciation followed by secondary contact with reductions of population size at 7,000 and 5,000 years ago respectively. I used RNA sequencing to investigate the gene expression divergence between Paxton Lake benthic and limnetic ecotypes and revealed genetic changes in *cis*-regulatory elements played an important role in the adaptation of benthic and limnetic ecotypes. Previous studies showed benthic and limnetic stickleback ecotypes from Enos Lake had been “collapsed” into a hybrid swarm due to the increased hybridization, whereas the genetic basis of this process is largely unknown. By investigating the whole genome re-sequencing data, I showed the “collapse” of Enos Lake species pair started earlier than previous prediction. Several genomic regions have been homogenized during the process, whilst others have not, which is possibly due to persistent divergent selection and/or low recombination rate at these regions.

Zusammenfassung

Sympatrische benthische (am Grund des Sees lebende) und limnische (im offenen Wasser lebende) Stichlinge entwickelten sich unabhängig voneinander in fünf Seen in Britisch-Kolumbien, Kanada. Da sie sich an unterschiedliche Nischen in ihrem Lebensraum anpassten, divergierten der benthische und limnische Stichlingsökotyp in ihrer Morphologie. Diese Evolution des benthischen und limnischen Stichlingsökotyps fand parallel in allen fünf Seen statt. Die Stichlinge dieser Seen bieten somit ein exzellentes Modell zur Untersuchung, welche Rolle die natürlicher Selektion bei der Speziation und der Anpassung spielt. Obwohl die Ökologie der Speziation und der Anpassung der benthischen und limnischen Stichlinge ausführlich untersucht wurde, fehlen bislang die genetischen Grundlagen dieser Mechanismen.

Ich verwendete Gesamt-Genom-Sequenzierung, um die Speziation und Anpassung von benthischen und limnischen Stichlingen in vier Seen (Paxton Lake, Priest Lake, Little Quarry Lake, Enos Lake) in Britisch-Kolumbien, Kanada, zu untersuchen. Benthische und limnische Stichlinge aller vier Seen zeigen parallele genetische Divergenz. Benthische und limnische Stichlingsökotypen waren stark divergierender natürlicher Selektion ausgesetzt, bei der abgeleitete und angestammte Allele in jeweils einer der Stichlingsökotypen selektiv favorisiert wurden. Im benthischen Ökotyp wurden erheblich mehr Genomregionen selektiert als im limnischen Ökotyp. Indem ich unterschiedliche statistische Ansätze kombinierte, identifizierte ich mit noch nie dagewesener Auflösung Genomregionen, die zur Anpassung des benthischen und limnischen Ökotyps beitragen. Dies ermöglicht mir die Identifizierung und Charakterisierung von Genen, die für die Anpassung der Ökotypen wichtige phänotypische Merkmale und biologische Prozesse kontrollieren. Durch die Verwendung von *high-density* genetischen Markern, die durch die Sequenzierung des gesamten Genoms generiert wurden, untersuchte ich die Abstammung der benthischen und limnischen Ökotypen und leitete daraus ein demographisches Modell für die benthischen und limnischen Stichlinge im Paxton Lake ab. Die benthischen und limnischen Stichlinge im Paxton Lake entstanden durch allopratrische Speziation gefolgt von sekundärem Kontakt, wobei die Populationsgröße jeweils vor 5.000 und 7.000 Jahren reduziert wurde. Ich verwendete RNA-Sequenzierung, um die Divergenz in der Genexpression zwischen dem benthischen und limnischen Ökotyp im Paxton Lake zu erforschen und deckte auf, dass genetische Veränderungen in *cis*-regulierenden Elementen eine wichtige Rolle in der Anpassung von benthischen und limnischen Ökotypen spielte. Bisherige Studien zeigten, dass benthische und limnische Stichlingsökotypen im Enos Lake auf Grund von erhöhter Hybridisierung in einen Hybridschwarm „kollabiert" waren. Die genetischen Grundlagen dieses Prozess sind jedoch größtenteils unbekannt. Durch Untersuchung der Gesamt-Genom-Sequenzierdaten zeigte ich, dass der Zusammenfall des Artenpaars im Enos Lake früher begann als bisher vorhergesagt wurde. Einige Genomregionen wurden bei diesem Prozess homogenisiert, andere nicht. Letzteres ist möglicherweise auf anhaltende divergente Selektion und/oder geringe Rekombinationsraten dieser Regionen zurückzuführen.

TABLE OF CONTENTS

1 GENERAL INTRODUCTION

Evolutionary biologists have been fascinated with studying speciation since Darwin first introduced the concept in his seminal book in 1859 (Darwin 1859). In this chapter, I first introduce the basic theories of population genetics and speciation. Secondly, I describe genetic approaches to identify regions under positive selection in the genome. Thirdly, I describe the recent advancements in adaptation genetics and genomics. Fourthly, I introduce the three-spined stickleback (*Gasterosteus aculeatus*) as an excellent model to study speciation and the recent advances in genetic and genomic studies of the sticklebacks. Lastly, I introduce benthic and limnetic sticklebacks, which are the species studied in this thesis, by describing their phenotypic traits and our knowledge of their speciation and adaptation to local environmental niches from ecological and genetic studies.

1.1 Population genetics

Population genetics pertains to the study of temporal and spatial changes of genetic variation in populations (Hedrick 2005). Early studies of genetic variation in natural populations predicted that only a limited number of genes would be variable [(Hedrick 2005), p295]. However, investigation of allozyme variation in human and *Drosophila pseudoobscura* populations found several polymorphisms, and individuals were often heterozygous at different loci, implying extensive variation of genes exists in a population (Harris 1966, Hubby & Lewontin 1966, Lewontin & Hubby 1966). These studies revolutionized population genetics, as this was the first time evolutionary geneticists quantified genetic variability more directly from studying proteins and not indirectly from morphological variations (Charlesworth & Charlesworth 2017). In addition, the discovery of extensive genetic variation in natural populations stimulated a new the debate on the role of natural selection and random genetic drift in maintaining variation in a population (Charlesworth & Charlesworth 2017).

Natural selection was considered the only force generating genetic variation in natural populations (Fisher 1958). Natural selection can be categorized into three main forms: positive selection, negative selection, and balancing selection (Hedrick 2005). Positive selection increases the frequency of alleles because they are beneficial to the survival and reproduction of individuals in the local environment. Negative (purifying) selection removes alleles that are deleterious or lethal. Balancing (stabilizing) selection maintains two or more alleles at one locus as heterozygotes have higher fitness than homozygotes (overdominance) or the fitness of alleles depends on its frequency (frequency-dependent selection). The scientists advocating the decisive role of natural selection consider balancing selection as the cause of genetic variation (Charlesworth & Charlesworth 2017, Gillespie 1991).

In contrast, Motto Kimura (1968) developed the “neutral theory of molecular evolution”, in which “genetic variation is primarily influenced by mutation generating variation and genetic drift eliminating it” (Kimura 1968)[(Hedrick 2005), p296]. The effective population size (N_e) is the number of breeding individuals in an ideal population in which “all parents have an equal expectation of being the parents of any progeny” [(Hedrick 2005), p205] and which maintains equal population size over generations (Hedrick 2005). Genetic drift describes the changes in allele frequency which are due to random sampling of gametes from generation to generation (Lynch et al 2011). Genetic drift reduces genetic variation from a population at a rate of $\frac{1}{2Ne}$, which is solely determined by N_e (Charlesworth 2009). The neutral theory proposed that the extensive genetic variation observed in natural populations resulted from genetic drift instead of natural selection because both negative selection and positive selection remove variation from populations much faster than genetic drift. Thus, the observed heterozygosity, which is the proportion of loci having two different alleles in a population, is the result of the equilibrium of effects of mutation and genetic drift [(Gillespie 2004), p29].

Although there was heated debate, there is a general consensus now that natural selection and genetic drift both shape the variation landscape of the genome (Hedrick 2005). Although scientists are usually more interested in loci influenced by natural selection than loci influenced solely by stochastic process (e.g., genetic drift), the neutral model is considered as the null model when detecting positive selection because stochastic processes are always occurring (Graur & Li 2000). To determine the relative contribution of selection and genetic drift to the frequency of alleles under directional selection, one can use the selection coefficient (s), the difference in fitness effects between alleles at the same locus (Nielsen 2005). The frequency of an allele is primarily determined by selection if $2N_es >> 1$, whereas the frequency is primarily determined by genetic drift if $2N_es << 1$ [(Gillespie 2004), p92], indicating genetic drift can randomly fix alleles, even the deleterious ones, in a population with a small population size. Furthermore, the strength of directional selection is positively correlated with N_e (Nielsen 2005). Therefore, elucidating the demographic history of populations is critical for not only better understanding their evolution, but also identifying regions under positive selection in the genome.

1.2 Statistical methods for detecting selection in the genome

Positive selection leaves a unique pattern of genetic variation at genomic regions influenced by it. As the neutral theory proposed that most of the genetic variation observed in a population arise from mutations and genetic drift, the signature of positive selection can be identified by the comparison to genome-wide background pattern of variations (Sabeti et al 2006). Therefore, the challenge of detecting regions under positive selection is to determine whether the pattern of genetic variation was derived from positive selection or random processes.

1.2.1 Methods for detecting selection based on changes of allele frequency

Several methods have been proposed to detect selection based on different types of population genetic data. The first type of data used to detect positive selection is genome-wide allele frequencies. Strong positive selection rapidly fixes the beneficial alleles in a population through a process called selective sweep (McVean 2007). In addition, neutral alleles that are closely linked with the selected alleles are fixed in the population faster than recombination can break the association between them. The fixation of neutral alleles closely linked to selected beneficial alleles is called genetic hitchhiking (Smith & Haigh 1974) (**Fig. 1.1b**). As positive selection increases the frequency of beneficial alleles, an excess of high frequency alleles can be observed in the target region, which can be detected by comparing the proportion of high frequency and intermediate frequency alleles in the region using Fay & Wu's H statistic (Fay & Wu 2000). After the selection pressure subsides, new mutations start to accumulate in the region, producing an excess of low frequency alleles. This excess of low frequency alleles is used in Tajima's D statistic to detect selection (Tajima 1989). However, one has to be cautious when using Tajima's D to detect positive selection, as both genetic hitchhiking and recent population expansion can generate an excess of low frequency alleles (Przeworski et al 2000, Tajima 1989).

Another feature of positive selection is the spatial pattern of genetic variation. As the selected beneficial alleles and linked neutral alleles are fixed during genetic hitchhiking, genetic diversity of the target region will be dramatically reduced in the population (Vitti et al 2013) (**Fig. 1.1a**). Therefore, a method was proposed to detect the signature of a selective sweep at a genomic region on the basis of deviations from a distribution of a simulated neutral allele frequency spectrum (AFS) (Kim & Nielsen 2004, Kim & Stephan 2002). Since Nielsen modified the method (composite likelihood ratio, CLR) to detect selective sweeps in genomic data using the AFS generated from empirical data (Nielsen et al 2005), researchers have detected several genomic regions under selection in different organisms (Long et al 2013, Pickrell et al 2009, Pool et al 2012).

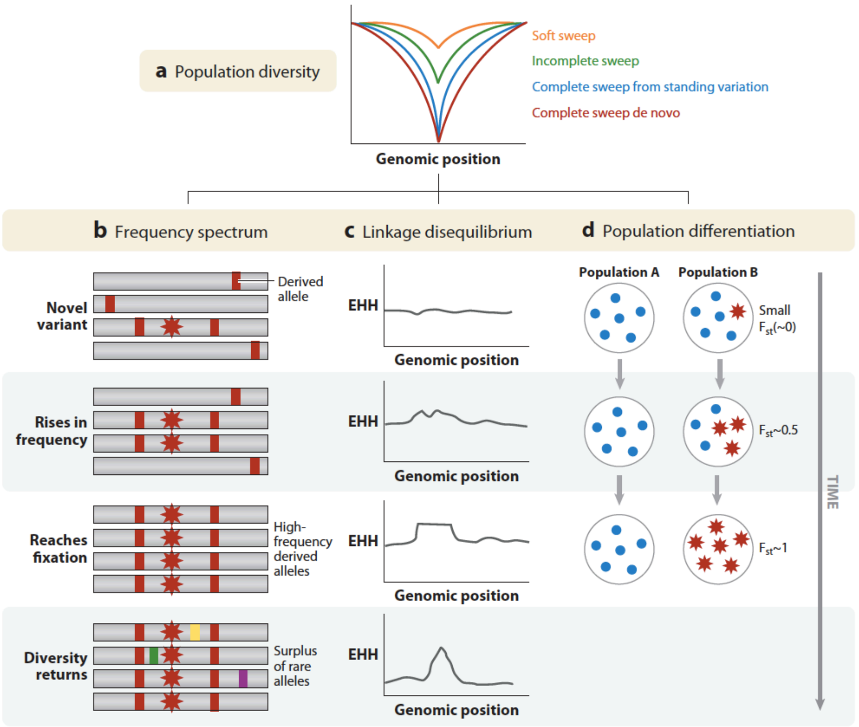

Figure 1.1 | Signatures of positive selection in population data. a, Different types of positive selection reduce genetic diversity of selected variants and linked neutral regions with differing intensities. In complete sweep of beneficial *de novo* mutation (hard selective sweep), positive selection rapidly drives the beneficial allele to fixation and increases the frequency of the linked neutral alleles, resulting in a sharp reduction of genetic diversity in target regions. In complete sweep from standing genetic variations (soft sweep with standing genetic variations), target beneficial pre-existing genetic variants are associated with different sets of neutral alleles due to historical recombination. Therefore, positive selection can reduce the genetic diversity of target alleles and a shorter region of neutral alleles. **b,** Positive selection shapes the site frequency spectrum of target region. A beneficial *de novo* mutation (red star) arises in the population. Positive selection increases its frequency and linked derived alleles (red bars), resulting in an excess of high-frequency derived alleles. After the beneficial allele fixes in the population and selection pressure subsides, new mutations (color bars) arise, resulting in an excess of low frequency alleles. **c,** Positive selection generates extended haplotype (set of genetic variants inherited together). Positive selection elevates the frequency of target alleles and linked neutral alleles quickly before recombination occurs in this region, generating extended homozygous haplotype. This can be detected by extended haplotype homozygosity (EHH) statistic. **d,** Positive selection increases genetic divergence between populations. Fixation index (F_{ST}) measures the level of differentiation between populations. Figure from (Vitti et al 2013).

1.2.2 Methods for detecting selection based on linkage disequilibrium

The second type of data that can be used to identify positive selection is linkage disequilibrium (LD). LD is a measure of association between two alleles on a chromosome [(Gillespie 2004), p101]. If the probability of two alleles being inherited together is high, these two alleles have high LD. Strong selection substantially increase the effect of genetic hitchhiking (Barton 2000). If the frequency of a beneficial allele increases rapidly enough, recombination does not have time to break down the linkage between the selected allele and nearby neutral alleles, resulting in a long haplotype (set of genetic variants inherited together) with a high frequency of homozygous alleles in the population (Sabeti et al 2002). The extended haplotype homozygosity (EHH) statistic was developed to detect highly homozygous haplotypes with high frequency in the population (Sabeti et al 2002) (**Fig. 1.1c**). The EHH measures the decay of homozygosity, as a function of distance, of haplotypes starting at a set of tightly linked variation sites ("core haplotype") to one end (**Fig. 1.2a**). To detect a signature of selection, the frequencies and EHH of different "core haplotypes" in one locus are compared. The core haplotype that has substantially higher frequency and EHH than other core haplotypes and simulated neutral sequences is considered to be under positive selection. As the original EHH test detects positive selection in a target locus, it is not suitable for identifying novel genomic regions under positive selection. Thus, the integrated haplotype score (iHS) method, which compares the extension of haplotypes carrying ancestral and derived core alleles, was developed for genomic scan of positive selection (Voight et al 2006) (**Fig. 1.2b**). To facilitate the genomic scan of positive selection, each iHS is normalized using empirical distribution of single nucleotide polymorphisms (SNPs) with the same derived allele frequency as the core allele. This test is able to differentiate between selection on *de novo* mutations or standing genetic variation by comparing the extension of haplotypes carrying derived and ancestral alleles.

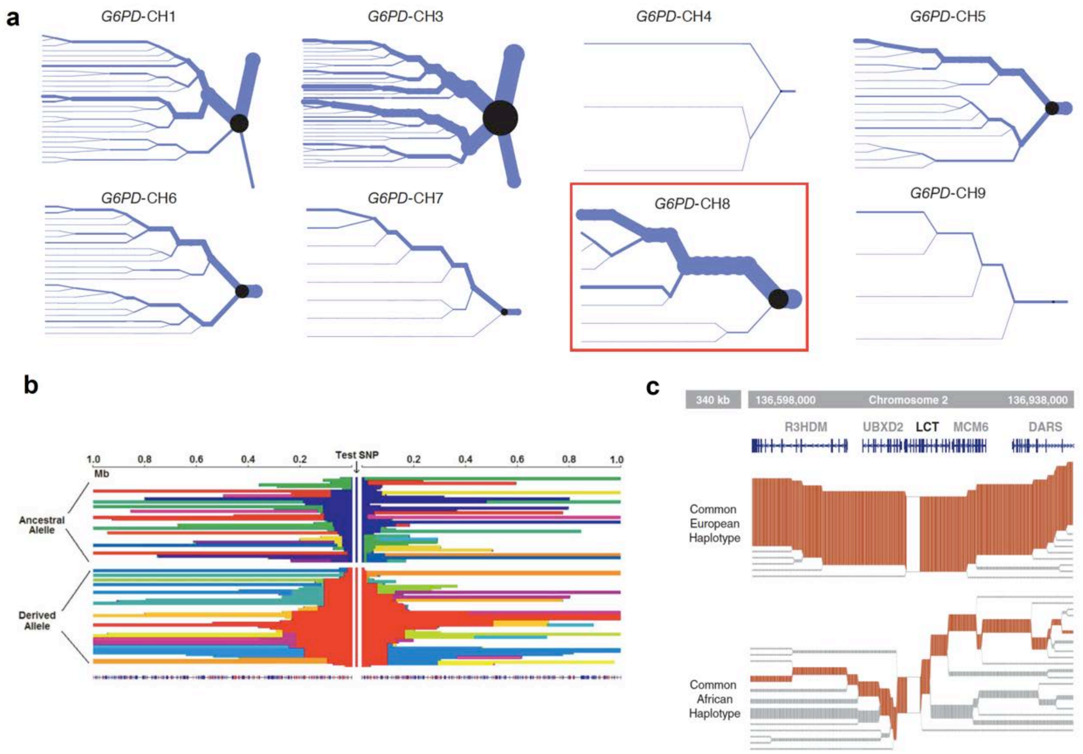

Figure 1.2 | Detecting positive selection based on the frequency and extension of homozygosity of haplotypes. a, The extension of homozygous haplotypes starting at different “core haplotypes” (indicated by black dots) at the *Glucose-6-phosphate dehydrogenase* (*G6PD*) locus, which is important for malaria resistance in humans. The haplotype *G6PD*-CH8 (red box) carrying the allele contributing to malaria resistance has both high frequency (denoted by the thickness of the line) and longer homozygous haplotype (the length of thick branch) than other core haplotypes in the African population, indicating a recent selective sweep at this allele. Figure from (Sabeti et al 2002) **b,** Extension of homozygous haplotypes carrying ancestral and derived alleles at a test SNP. Homozygosity of haplotypes carrying the derived allele have higher frequency and extend longer than the haplotypes carrying the ancestral allele, indicating the derived allele at this site underwent recent positive selection. **c**, Haplotypes carrying the lactase (LCT) persistence allele in European and African populations. The haplotype carrying the lactase persistence allele (indicated by orange lines) in the European population has high frequency and homozygousity, suggesting a recent selective sweep in the European population. On the other hand, the haplotype carrying the allele is not common in the African population. Figure from (Sabeti et al 2006).

1.2.3 Methods for detecting selection based on population differentiation

The third type of data that can be used to detect positive selection is population differentiation. Nearly all species have several populations with varying degrees of isolation (Holsinger & Weir 2009a). These populations usually live in different environmental niches and are subject to different environmental pressures. Therefore, phenotypic traits that contribute to local adaptation of populations residing in divergent environments might be different (Vitti et al 2013). If selection acted on one population but not the other, allele frequencies at the selected locus and nearby neutral sites between these two populations can differ substantially (**Fig. 1.1d**) (Vitti et al 2013). On the other hand, genetic differentiation at neutral regions is mainly determined by genetic drift. Genetic drift can remove or fix alleles at neutral regions over time, but requires significantly longer times than selection (Holsinger & Weir 2009a). Therefore, genomic regions with significantly higher genetic differentiation than the genome-wide level are considered to have been subject to selection (Vitti et al 2013). After Sewell Wright introducing the concept in 1931, the fixation index (F_{ST}) has become the most commonly used measure of genetic differentiation among populations (Holsinger & Weir 2009a). However, genetic differentiation at neutral regions is determined by genetic drift, and the effect of drift is highly variable in the genome. In addition, as F_{ST} is a single nucleotide measurement, it is possible that one (or more) neutral site possesses high F_{ST} by chance, making it difficult to distinguish selective regions from neutral regions that are highly differentiated between populations (Chen et al 2010). Therefore, several statistical tests which integrate genetic differentiation with other statistics have been developed to improve the power of selection detection (Vitti et al 2013). First, cross-population extended haplotype homozygosity (XP-EHH) was developed to detect positive selection by comparing EHH of core alleles in two populations (**Fig. 1.2c**) (Sabeti et al 2006). Second, cross-population composite likelihood ratio test (XP-CLR) identifies the signature of selection by calculating the composite likelihood of deviation of allele frequency differentiation to neutral expectation across multiple variation sites (Chen et al 2010). Both XP-EHH and XP-CLR tests utilize the idea that genetic hitchhiking affects large flanking regions, resulting in either extended LD (XP-EHH) or extended regions of low diversity (XP-CLR), while genetic drift can only increase genetic differentiation of unlinked neutral sites.

Different methods can detect selection that occurred at various times in history because these methods identify different signatures left by selection (Sabeti et al 2006) (**Fig. 1.3**). For example, an excess of high frequency derived alleles or low frequency alleles can only be detected when the target beneficial allele is fixed or after it is fixed in the population. Thus, statistical tests that detect selection based on shifts in the allele frequency spectrum can detect selection that occurred a long time ago. On the other hand, statistical tests based on the length of haplotypes detect unusual extension of homozygous haplotypes before recombination breaks down the linkage, and are thus suitable to identify signature of ongoing selection. As a result, to obtain a comprehensive genomic landscape of selection, tests have been developed to detect positive selection based on different signatures of selection. Methods of detecting positive selection by calculating composite probability of different tests have been shown to detect more regions under selection with higher accuracy and resolution in humans (Grossman et al 2013, Grossman et al 2010, Pickrell et al 2009).

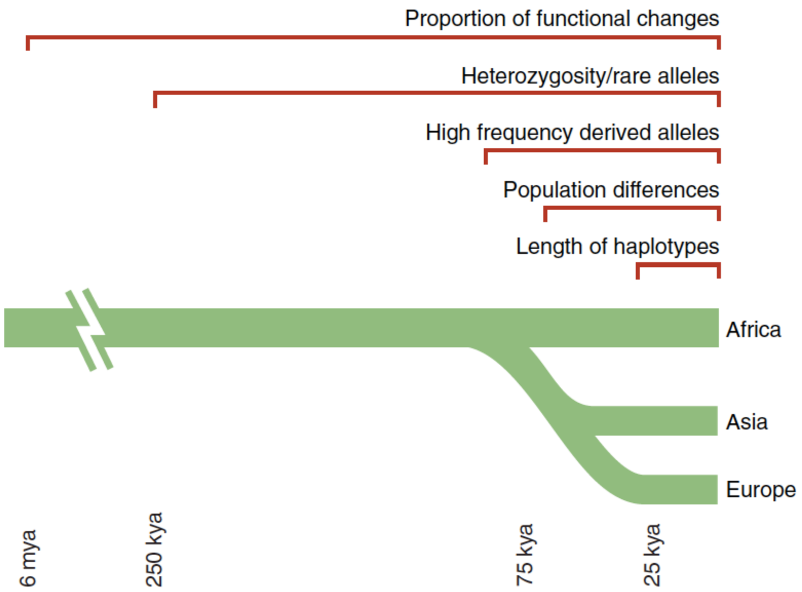
Proportion of functional changes
Heterozygosity/rare alleles
High frequency derived alleles
Population differences
Length of haplotypes
Africa
Asia
Europe
6 mya
250 kya
75 kya
25 kya

Figure 1.3 | Signatures of selection occurred at different historical time in humans. Methods based on different signatures can detect selection that occurred at different times in history. Figure from (Sabeti et al 2006)

1.2.4 Challenges of detecting positive selection

Although several genomic loci under positive selection have been successfully identified in diverse organisms using the methods described above, there are still several challenges of identifying regions subject to positive selection in the genome. First, one needs to distinguish genetic hitchhiking from background selection, which is the process purifying selection that eliminates recurrent deleterious alleles generated by mutation and linked neutral variants in regions with low recombination (Charlesworth et al 1993, Nordborg et al 1996). Background selection can reduce the local effective population size (N_e), which further reduces the genetic diversity of affected regions, mimicking the pattern of genetic hitchhiking (Charlesworth et al 1993, Stephan et al 1999). A study of background selection in regions with normal recombination rates showed that background selection is unlikely to generate large genomic regions of reduced diversity in these regions (Loewe & Charlesworth 2007). Therefore, statistical tests have been developed to identify genomic regions under positive selection by taking local recombination rate into account (DeGiorgio et al 2016).

Second, selection on pre-existing standing genetic variation creates a different pattern of genetic variation compared to genetic hitchhiking described above (Hermisson & Pennings 2005, Przeworski et al 2005). In genetic hitchhiking, a new beneficial mutation sweeps through the population, resulting in a skewed allele frequency spectrum, extension of homozygous haplotypes, and strong reduction of local genetic diversity (Jensen 2014). In contrast, some neutral or nearly neutral alleles maintained in the population by genetic drift can become beneficial if the environment changes, and positive selection can act on these pre-existing variants and drive them to fixation quickly (Hermisson & Pennings 2005). As these standing genetic variants segregate in the population for a long time, they can associate with different haplotypes due to recombination before the selection shift (Przeworski et al 2005) (**Fig. 1.4**). Thus, the sweep of beneficial standing genetic variants would carry diverse haplotypes to intermediate frequency, resulting in a moderate reduction of genetic diversity (**Fig. 1.1a**). To differentiate these two types of sweeps, the selective sweep of new mutations is termed the classical hard selective sweep, while a sweep of standing genetic variants is called a soft selective sweep (Hermisson & Pennings 2005). As soft sweeps generate different signatures of selection compared to hard sweeps, most of the previously described methods that were developed for detecting genetic hitchhiking are not able to detect soft sweeps (Vitti et al 2013). A study on simulated sequences showed that methods based on allele frequency changes are unable to detect soft sweeps, and methods based on the extension of haplotypes have reduced power to detect soft sweeps (Pennings & Hermisson 2006). Because theoretical and functional analyses showed that selection on standing genetic variation is important for adaptation and pervasive in the genome (Messer & Petrov 2013, Pritchard & Di Rienzo 2010, Wilson et al 2017), methods that are specifically designed for detecting soft sweeps in the genome have been proposed recently (Garud et al 2015, Peter et al 2012, Schrider & Kern 2016).

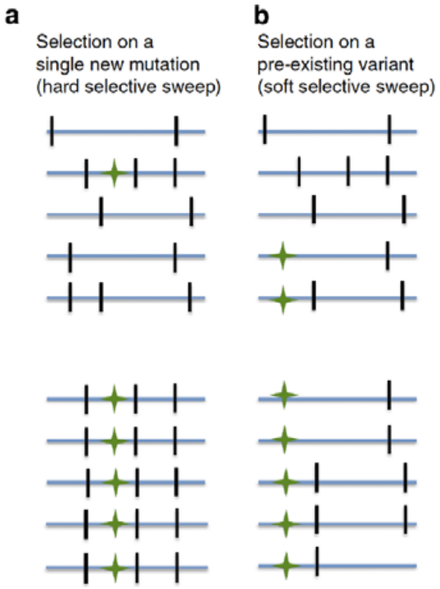

Figure 1.4 | Signatures of selection on *de novo* (new) mutations and standing genetic variation. a, Change in patterns of genetic variation before selection (top) and after selection (bottom) during a hard sweep. A new beneficial allele arises in one individual (green star, top panel) and rapidly sweeps through the population by positive selection (bottom panel), carrying several neutral alleles (black bars) with them. **b,** Change in patterns of genetic variation before selection (top) and after selection (bottom) during selection on standing genetic variation. Pre-existing genetic variants (green stars, top panel) become beneficial and quickly sweep through the population (bottom panel), carrying two distinct haplotypes with them. Figure modified from (Jensen 2014)

1.3 Speciation

The study of how species evolved from populations (speciation) is one of the most important subjects of evolutionary biology (Coyne & Orr 2004). Speciation is the research subject that connects the study of continuous genetic variations in populations that I described in the previous section (microevolution) and the study of diverse discrete species in the nature (macroevolution) (Weissing et al 2011). Therefore, studying speciation help us to understand how changes of genetic variations in populations result in the huge biodiversity observed in nature.

1.3.1 Reproductive isolation

After the formal introduction of reproductive isolation as the definition of species by Dobzhansky and the pioneering empirical works by Dobzhansky (Dobzhansky 1936) and Muller (Muller & Pontecorvo 1942), researchers started to gain knowledge about speciation (Coyne & Orr 2004, Orr 2001, Seehausen et al 2014). Most evolutionary biologists have since adopted the biological species concept that was first proposed by Mayr, which defines species as “interbreeding natural populations that are reproductively isolated from other such groups” (Mayr 1942). Thus, speciation is the emergence and preservation of reproductive barriers between populations that ensure the maintenance of genetic and phenotypic divergence (Coyne & Orr 2004, Seehausen et al 2014). As reproductive isolation is the essence of the definition of species, understanding reproductive isolation between species is considered a major subject in the study of speciation (Coyne & Orr 2004). Mechanisms of reproductive isolation can be classified as extrinsic or intrinsic factors.

Individuals from populations living in distinct environments might develop morphological traits adapted to their local habitats. As a result, immigrants may suffer lower viability or reproductive success than the resident population, which is called extrinsic prezygotic isolation (Schluter & Conte 2009). Even after hybrids are produced, hybrids may suffer lower viability or reproductive success in both parental environments if they have intermediate phenotypes (Coyne & Orr 2004, Schluter 2009). This is termed extrinsic postzygotic isolation.

Other mechanisms of reproductive isolation are classified as intrinsic reproductive isolation, as they do not require interaction with the environment (Coyne & Orr 2004). For example, assortative mating, in which females are more likely to mate with males having similar phenotypic traits, is classified as intrinsic prezygotic isolation. Lastly, in intrinsic postzygotic isolation, hybrids are inviable or sterile due to developmental defects caused by genetic properties of the individuals. The widely accepted genetic model of intrinsic postzygotic isolation is the Bateson-Dobzhansky-Muller incompatibility (BDMI) (Bateson 1909, Dobzhansky 1936, Muller 1942). According to the BDMI model, derived alleles are fixed in different loci in two populations separately. Although the derived alleles are not deleterious in their own genomic background, the negative epistatic interactions cause negative effects when these two alleles bring together through hybridization.

1.3.2 Geographic model of speciation

Darwin considered natural selection plays critical role in the origination of species (Darwin 1859). However, due to the limited knowledge of inheritance, Darwin only provided verbal arguments of the role of natural selection in speciation. In addition, as theoretical studies showed speciation by natural selection was unlikely, Mayr emphasized the role of geographic isolation of populations in the origination of species (geographic model of speciation) (Weissing et al 2011).

In geographic model of speciation, speciation can be classified as allopatric speciation, parapatric speciation, or sympatric speciation according to the degree of geographic separation and extent of gene flow between diverging populations (Coyne & Orr 2004). Allopatric speciation is the emergence of new species from populations where mating is not possible between the subpopulations because of geographical isolation (Gavrilets 2003). Sympatric speciation occurs under random mating between incipient subpopulations occupying same environment during speciation (Gavrilets

2003, Mayr 1963). Parapatric speciation is a model in which subgroups of population adapted to continuous environmental niches genetically diverge and reduce migration and mating, and finally become independent species (Gavrilets 2003).

The prevalence of sympatric and allopatric speciation is one of the most controversial questions in the study of evolution (Coyne & Orr 2004). Because of Mayr's famous critique of sympatric speciation, which claimed interbreeding and recombination would rapidly break down the linkage of gene complexes contributing reproductive isolation, some evolutionary biologists expected sympatric speciation to be uncommon in nature (Coyne & Orr 2004). Therefore, allopatric speciation was the main topic of speciation studies in the past (Coyne & Orr 2004). Theoretical studies proposed three main stages of allopatric speciation: first, an ancestral population splits into isolated populations due to a sudden geographic change or colonization of a novel habitat; second, genetic divergence between isolated populations arise because divergent selection and genetic drift fix different alleles in these populations; third, genetic divergence produces reproductive isolation when isolated populations experience secondary contact and they reside in sympatry thereafter (sexual selection can reinforce the isolation by limiting interbreeding) (Coyne & Orr 2004). Researchers have identified numerous examples of allopatric speciation in nature (Lowry et al 2008, Sobel et al 2010).

Sympatric speciation started to gain the attention of evolutionary biologists since the 1990s partly due to the development of molecular phylogenetics and studies of the enormous diversity of sympatric cichlid fish in different African lakes (Bolnick & Fitzpatrick 2007, Coyne & Orr 2004). People proposed two interacting models of sympatric speciation: character displacement, in which reproductive isolation arises from disruptive natural selection involving competition for resources; and disruptive sexual selection, in which female preference drives differentiation of male traits (Bolnick & Fitzpatrick 2007, Schluter 2000). Disruptive natural selection is considered as a major cause of sympatric speciation (Coyne 2007, Schluter 2001). If the genomic loci under disruptive natural selection are linked with loci causing assortative mating, disruptive natural selection can initiate assortative mating and sexual selection between diverging species. In the end, disruptive natural selection and sexual selection reinforce each other and generate reproductive isolation (van Doorn et al 2009). Other selection pressures, such as sexual conflict and male-male competition, were also shown to initiate assotative mating and interact with sexual selection to form reproductive isolation during sympatric speciation (Bolnick & Fitzpatrick 2007).

Although difficult, scientists have found several empirical examples of sympatric speciation. The most convincing example of sympatric speciation in nature is the African cichlid fish. Scientists found that the diverse cichlid fish species from different African crater lakes evolved from sympatric speciation based on phylogenetic and population genomic analyses (Barluenga et al 2006, Malinsky et al 2015, Meyer et al 1990, Schliewen et al 1994). Cases of sympatric speciation were also found in other fish species and plants (Crow et al 2010, Gislason et al 1999). Therefore, theoretical and empirical studies have demonstrated that sympatric speciation is feasible, even if it is not common in nature.

1.3.3 Ecological speciation

The geographic model of speciation classifies speciation based on the geographic separation of populations, which does not facilitate the study of evolutionary mechanisms driving the generation of reproductive isolation (Schluter 1998). Therefore, classification according to the evolutionary mechanisms has been proposed, which classified speciation into speciation by nature selection, speciation by drift, and polyploidy speciation (Schluter 2001). Recent advances of speciation research demonstrated speciation by natural selection was common in nature (Schluter 2009, Schluter & Conte 2009, Weissing et al 2011). According to the degree of involvement of ecological factors in the process, speciation by natural selection can be classified as mutation-order speciation or ecological speciation (Schluter & Conte 2009). Mutation-order speciation is the process of fixing beneficial but incompatible mutations in different populations under similar selective pressure (Schluter 2009). Ecological speciation, which is the process where reproductive isolation arises from ecologically divergent natural selection during adaptation of populations to contrasting environments, is one of the most important subjects of speciation research (Dieckmann et al 2004, Rundle & Nosil 2005).

The genetic basis of prezygotic and postzygotic isolation in ecological speciation has been studied extensively (Schluter & Conte 2009). Immigrant inviability and assotative mating are two major causes of prezygotic isolation in ecological speciation (Nosil et al 2005, Schluter & Conte 2009). The degree of immigrant inviability increases as divergent natural selection drives populations to their fitness optimum (Nosil et al 2009b). In ecological speciation, assortative mating can arise from the process in which females distinguish conspecific males according to phenotypic traits regulated by loci under divergent selection (Felsenstein 1981). In addition, natural selection might increase the divergence of adaptive loci and the tightly-linked loci contributing to assortative mating in regions with low recombination, which promotes assortative mating between populations (Schluter & Conte 2009). Divergent selection can also generate postzygotic isolation between populations. As natural selection drives the adaptation of populations to diverging environments, hybrids suffer from reduced fitness in both parental ecological niches due to their intermediate phenotypes (Rundle & Whitlock 2001, Schluter & Conte 2009).

1.4 Adaptation genetics and genomics

A major challenge in evolutionary biology is to elucidate the relative contribution of stochastic processes (i.e. genetic drift) and natural selection in the species origination and diversification (Elmer & Meyer 2011). The ecological speciation model described in previous section demonstrates adaptation to contrasting environments through natural selection can generate reproductive isolation between populations. In addition, studies revealed the prominent role of natural selection in generating morphological diversification in closely related groups within species during adaptation (Berner & Salzburger 2015, Elmer & Meyer 2011). Therefore, investigating the genetic and genomic basis of adaptation provides valuable insights of how biodiversity originated in nature. Evolutionary biologists have made great progress in the study of adaptation by identifying adaptive loci and genomic patterns of divergence in different organisms (Berner & Salzburger 2015, Savolainen et al 2013).

1.4.1 Molecular mechanism of adaptation

1.4.1.1 Genetic basis Adaptive loci

Identifiying and charaterizing adaptive loci is one of the most important subject in the study of adaptation. Evolutionary biologists historically believed adaptation involved mutations at multiple loci with small effects (Orr 2005). Therefore, it is impossible to identify and chacterize genes contributed to adaptation (adaptive loci) as the number is too large. However, recent efforts using genetic mapping identified several genes that can explain large portion of phenotypic variation (effect size) of traits contributing to the adaptation of different populations/species (Pardo-Diaz et al 2015). For example, *melanocortin-1 receptor* (*Mc1r*) and *Agouti* loci control the coat color transition from dark in mainland mice to light beach mice (Hoekstra et al 2006, Manceau et al 2011). Whereas, QTL mapping studies also found adaptive phenotypic changes can be regulated by several loci with small effects (Orr 2005).

The identification of adaptive loci in diverse species also enable evolutionary biologists to investigate another important question in the study of adaptation: whether parallel phenotypic adaptation involve the same set of genomic loci (Elmer & Meyer 2011). Genetic studies demonstrated the same gene could regulate the transition of traits in divergent populations adapted to similar environments. For example, repeat reduction of armor plates in sticklebacks during adaptation from marine to freshwater environment is largely caused by mutations in Ectodysplasin (*Eda*) locus (Colosimo et al 2005, Colosimo et al 2004). However, adaptation to similar environments does not necessarily require selection on the same gene, even in closely related populations of the same species. A study showed that *Mc1r* controlled coat color transition in populations of rock pocket mice from a region in

Arizona, USA, but not population from the nearby region in New Mexico, USA, indicating that anther locus (or loci) should regulate coat color in populations in New Mexico (Hoekstra & Nachman 2003, Nachman et al 2003). This suggests the genetic basis of adaptation is complicated, and our understanding of adaptation is far from articulating theories or making predictions (Elmer & Meyer 2011). Thus, more genomic regions contributed to populations' adaptation need to be identified.

Although identifying and analyzing adaptive loci in various organisms have provided insight into how natural selection shape traits during adaptation, genetic mapping of adaptive loci have several limitations: 1) hybrids between studying populations must be viable and reproducible, which is impossible in some species, 2) it is limited to adaptive traits that are easy to dissect, 3) it is confined to loci with large effects due to technical limitation (Savolainen et al 2013). Theoretical and empirical studies suggest there are more loci with small effects than loci with large effects that contribute to adaptation (Orr 2005). Thus, it is critical to switch from identifying single adaptive loci with large effects to comprehensive genomic scans of adaptive loci. With the advent of next-generation sequencing and the development of statistical methods of detecting genomic regions under natural selection described in **Section 1.2**, evolutionary biologists have successfully identified several adaptive loci in diverse species (Berner & Salzburger 2015).

1.4.1.2 Contribution of coding and regulatory changes in adaptation

One of the important insights of adaptation scientists learned from charactering adaptive loci is adaptation can be achieved by genetic changes at both coding and regulatory sequences. Identifying and characterizing adaptive loci demonstrated coding changes contribute to adaptive morphological changes in several species (Hoekstra et al 2006, Protas et al 2006, Werner et al 2005a, Werner et al 2005b). In contrast, genetic mapping and analysis of genomic loci controlling adaptive morphological modifications found that changes in regulatory sequences contributed to adaptation in numerous species (Jeong et al 2008, Martin et al 2012, Rebeiz et al 2009, Reed et al 2011, Wray 2007b). Thus, the relative contribution of coding and regulatory changes in speciation has been under considerable debate (Hoekstra & Coyne 2007, Wray 2007b).

The early approaches of population genetics were restricted to studying coding sequence variations in natural populations due to the limitation of knowledge and methodology (Wray 2007b). Evolutionary biologists developed several theoretical models explaining the role of coding changes in speciation and adaptation. In addition, genetic and genomic studies in diverse organisms showed coding sequence variations of adaptive loci contributing to their speciation and adaptation (Hoekstra & Coyne 2007).

The contrasting hypothesis suggests that modifications of gene expression by changes in regulatory regions play a prominent role in evolution and adaptation (Carroll 2008, Wray 2007b). This hypothesis suggests that phenotypic evolution of organisms is largely due to changes in regulation of gene expression of functionally-conserved proteins through mutations in *cis*-regulatory elements that control expression of a single nearby gene, or *tran*-regulatory factors that regulate expression of several downstream genes elsewhere in the genome (Carroll 2008, Stern & Orgogozo 2009). A single gene can have multiple *cis*-regulatory elements (e.g., promoters and enhancers) that serve as binding sites for *trans*-regulatory factors (i.e., transcription factors) (Mack & Nachman 2017). These interacting *cis*- and *trans*-regulatory elements regulate the expression of the target gene (Stern & Orgogozo 2009). Both theoretical and empirical studies of gene expression regulation have demonstrated that the divergence in *cis*- or *trans*-regulatory sequences (*cis*- or *trans*- regulatory divergence) contributes to adaptation (Jones et al 2012b, Stern & Orgogozo 2009).

Investigating single adaptive loci is not sufficient to evaluate the relative contribution of coding and regulatory changes to adaptation, as it is biased toward loci with large effect size. Therefore, it is critical to apply genomic approaches to comprehensively investigate the relative importance of these two mechanisms. For example, Pollard et al. (2006) compared available animal reference sequences and found almost all (96%) genomic regions with significantly accelerated rates of substitutions in humans were located in regulatory regions (Pollard et al 2006). However, most of the genomic studies of this subject do not consider the phenotype and thus neglect the fact that some of these regulatory changes influence the expression of the genes that do not contribute to adaptation of a population. Thus, it is of great important to study the relative contribution of coding and regulatory changes to adaptation using approaches combining comparative genomics and expression divergence analysis.

1.4.2 Evolutionary processes of adaptation

1.4.2.1 Genetic architecture of adaptation

Describing the number and distribution of adaptive loci in the genome is of great importance and has become one of the most active areas in speciation research (Noor & Feder 2006). In contrast to the hypothesis that only a few genomic loci with large effects promote adaptation, numerous genomic regions were found to be involved in adaptation (Seehausen et al 2014). A recent review of published genomic studies of various species found that 5-10% of genomic loci were shaped by disruptive natural selection and highly diverged between populations (Nosil et al 2009a). These highly divergent regions were distributed on different chromosomes and dispersed on the background of low divergence (Nosil et al 2009a). Divergent natural selection is considered to play a prominent role in generating this genomic pattern of heterogeneous divergence (Nosil et al 2009a). The divergence of closely linked neutral genomic regions of adaptive loci is expected to increase due to the effect of genetic hitchhiking. In addition, gene flow between sympatric species or allopatric species experiencing secondary contact reduces the divergence of other regions and creates backgrounds of low divergence (Nosil et al 2009a, Via 2009). This selection-with-gene-flow model can further generate large “island of genomic divergence” (Feder et al 2012, Via 2012). First, the divergent genomic regions extend due to genetic hitchhiking. Second, hybridization at these extended regions cause hybrids to suffer lower fitness. Thus, gene flow and local recombination are reduced at these regions, allowing some of them with close genetic distances to form the “island of divergence” (“divergent hitchhiking”).

Genomic studies of divergence landscapes have found these “islands of divergence” in several species, including *Heliconius* butterflies, Darwin’s finches, *Ficedula* flycatchers, Atlantic cod, sunflowers, crows, house mice, and African malaria mosquitoes (Alonso-Blanco et al 2016, Brawand et al 2014, Ellegren et al 2012, Harr 2006b, Hemmer-Hansen et al 2013, Lamichhaney et al 2015, Nadeau et al 2012, Poelstra et al 2014, Renaut et al 2013, Turner et al 2005, White et al 2010). However, the “island of divergence” is not a universal phenomenon. The highly divergent genomic regions can be not clustered but distributed on different chromosomes in other species (Brawand et al 2014, Harr 2006a). Linkage between locally adapted alleles could promote adaptation of populations (Kirkpatrick & Barton 2006, Nachman & Payseur 2012). In contrast, strong linkage between adaptive and maladaptive loci can deleterious, which impedes adaptation (Barton 2010). As evolutionary biologists just started to obtain knowledge of genomic architecture of adaptation using genomic approaches, it is critical to investigate the adaptive landscape in natural populations and provide empirical evidences to this question.

1.4.2.2 Source of adaptive variation

The initial genetic variation in adaptive loci is considered to originate primarily from *de novo* mutations and standing genetic variation (Hedrick 2013). Owing of the assumptions of natural selection used for detecting selective sweeps in different statistical programs, most of the adaptive loci identified so far using population genomic approaches are thought to have originated from *de novo* mutations (Przeworski et al 2005). However, current theoretical and empirical studies indicate that adaptation from standing variation is of great importance (Barrett & Schluter 2008a, Garud et al 2015, Hermisson & Pennings 2005, Messer & Petrov 2013, Reid et al 2016). Taken together, it is crucial to identify the origination of genetic variation from these two sources. Thus, analyses differentiation both selections on *de novo* mutation and standing variation would provide a more general idea of how adaptive variation originate.

1.5 Threespine stickleback fish

1.5.1 The threespine stickleback is a good model to study adaptation

The threespine stickleback (*Gasterosteus aculeatus*) is a species complex comprising thousands of phenotypically diverse populations, and serves as an excellent model to study adaptation (Bell & Foster 1994b, McKinnon & Rundle 2002). Marine sticklebacks started to invade diverse freshwater systems in the northern hemisphere about 12,000 years ago after the last glacial retreat (McPhail 1993). During this short period of time, freshwater sticklebacks have evolved into many ecotypes adapted to different environments (Bell & Foster 1994b).

Different freshwater stickleback populations evolved similar traits recurrently during colonization of similar freshwater environments (McKinnon & Rundle 2002). Repeated and independent evolution of traits in association with environmental variables rather than spatial distance is one of the powerful features of the stickleback system and has been studied in depth for numerous traits (Bell & Foster 1994b). There are numerous phenotypic variations between marine and freshwater sticklebacks, including armor plate number, presence/absence of pelvic spine and dorsal spine, body size, body shape, body color, and courtship behavior (Bell & Foster 1994b). Armor plate number, presence/absence of pelvic spine and dorsal spine, and body size are the most discriminating characters between marine and freshwater sticklebacks (**Fig. 1.5**) (Reimchen et al 1985). Unlike most of the fishes possessing scales, sticklebacks have special body armor comprised of bony lateral plates, dorsal spines, and a spined pelvic girdle, which help stickleback escape from predation (Bell & Foster 1994b, Reimchen 1994). Because of the higher growth cost of mineralizing bone in low ion concentration environments and the reduction of predators, freshwater sticklebacks lost armor plates and pelvic spines during their adaptation (Spence et al 2013, Spence et al 2012). As a result, marine sticklebacks usually have a complete row of armor plates covering head to tail ("complete" morph), while freshwater sticklebacks have partial or no armor plates covering the body ("partial" or "low" morph) (Bell & Foster 1994b). Taken together, these observations suggest natural selection plays an important role in generating the morphological variations in sticklebacks (Berner & Salzburger 2015).

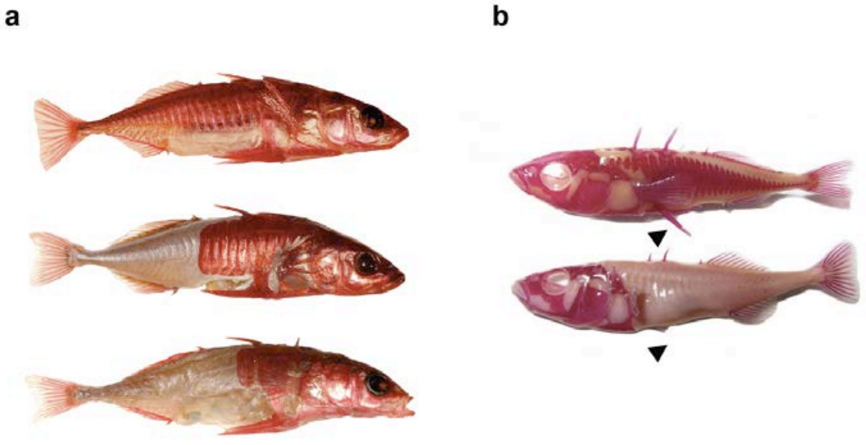

Figure 1.5 | Morphological divergence of sticklebacks. **a,** From top to bottom, the “complete”, “partial”, and “low” morph of armor plates of sticklebacks. To better illustrate the armor plates, fishes were stained with Alizarin red. Figure from (Barrett et al 2008) **b,** Sticklebacks with (top) and without pelvic spines (bottom), the black arrows point out the pelvic spine of sticklebacks. Figure from (Cleves et al 2014).

After a certain level of reproductive isolation, populations start to accumulate their own genetic variation due to mutations, genetic drift, and selection, which lead to further reproductive isolation (Nosil et al 2009b). Studying different stages of reproductive isolation provides valuable insights into the mechanisms of speciation (Seehausen et al 2014). The stickleback is a good system to study speciation because different stickleback population pairs have diverse strengths of reproductive isolation with the genetic differentiation between populations measuring by Nei’s D ranging from low in lake-stream pairs (very low) to medium in marine-freshwater pairs (0.008) to high in Japanese species pairs (0.428) (McKinnon & Rundle 2002).

Existing powerful genetic and genomic tools also make sticklebacks a good system to study adaptation and speciation. The fact that hybrids of ancestral (marine) and derived (freshwater) individuals are viable enables researchers to map adaptive loci in sticklebacks (Kingsley & Peichel 2007). Moreover, a high quality genetic map (Peichel et al 2001), a reference sequence (Jones et al 2012a), genome-wide resequencing datasets (Jones et al 2012a, Jones et al 2012b, Marques et al 2016, Roesti et al 2015), BAC libraries (Kingsley et al 2004), transgenic methods (Tol2 (Chan et al 2010) and CRISPR/Cas9 (Hart & Miller 2017)) and a mature microinjection protocol (Erickson et al 2016) exist for this model, enabling excellent studies of stickleback adaptation and speciation.

1.5.2 Adaptive genetics and genomics of sticklebacks

Scientists have successfully cloned and studied the function of several adaptive loci in sticklebacks. Reduction of armor plate number is one of the major changes during stickleback adaptation to freshwater environments. The gene controlling armor plate number has been mapped and studied intensively. A major QTL and several other QTLs with small effect controlling armor plate number were identified in sticklebacks using genetic mapping (Colosimo et al 2004, Cresko et al 2004). *Eda* locus was later identified as the major QTL controlling repeat reduction of armor plate number in sticklebacks (Colosimo et al 2005). Genetic changes in the enhancer of the *Eda* locus have been found to cause the reduction of armor plates in freshwater sticklebacks (O'Brown et al 2015). The low-plate *Eda* allele has been repeatedly selected during the adaptation of freshwater sticklebacks due to the faster growth rate of low plated fishes in water of low ion concentration (Barrett et al 2008, Colosimo et al 2005, Raeymaekers et al 2014, Schluter et al 2010). Pelvic spine reduction is another major morphological change during freshwater stickleback adaptation (Reimchen & Nosil 2006). Repeated *de novo* deletions in the enhancer region of the *Pituitary homeobox transcription factor 1* (*Pitx1*) gene have caused pelvic reduction in different freshwater stickleback populations (Chan et al 2010, Shapiro et al 2006). In addition, *cis*-regulatory changes in Kit ligand (*Kitlg*) and Growth/Differentiation Factor 6 (*GDF6*) have been shown to contribute to the changes in gill/ventrum pigmentation and armor plate size in freshwater sticklebacks (Indjeian et al 2016, Miller et al 2007). Furthermore, *cis*-regulatory change of the *Bone morphogenetic protein 6* (*Bmp6*) gene was discovered to result in gain of the ventral pharyngeal tooth in freshwater sticklebacks (Cleves et al 2014).

Mapping and dissecting adaptive loci in sticklebacks has greatly improved our understanding of their adaptation. First, genetic changes controlling adaptive traits studied in sticklebacks are caused by mutations in regulatory sequences of genes, indicating an important role of regulatory changes in stickleback adaptation. This might due to the fact that each of these adaptive genes regulates several different developmental processes (pleiotropy), genetic changes in the coding sequence of the gene might have deleterious pleiotropic effects. Spatial expression regulation of these genes in a particular developmental process can generate the morphological divergence among different populations. Second, adaptive variations can be derived from both *de novo* mutations and standing genetic variation. The alleles controlling the reduction of armor plates and transition of gill/ventrum pigmentation in freshwater sticklebacks were found at low frequency in marine sticklebacks, suggesting selection for standing genetic variants contributed to these two morphological transitions (Colosimo et al 2005, Miller et al 2007). Conversely, repeated reduction of pelvic spine in diverse freshwater sticklebacks is due to recurrent *de novo* deletions in the enhancer of *Pitx1* gene (Chan et al 2010). Genomic study of global marine and freshwater sticklebacks demonstrated the prominent role of reusing standing genetic variations during freshwater sticklebacks adaptation (Jones et al 2012b).

Genomic study of speciation and adaptation in stickleback is feasible due to the relatively small genome size (463 Mb) and high-quality reference sequence assembly (Jones et al 2012a). Using genome-wide variation datasets, numerous highly divergent loci have been identified between marine and freshwater stickleback populations, as well as freshwater populations separated by different geographic distances (Deagle et al 2012, Ferchaud & Hansen 2016, Hohenlohe et al 2010, Jones et al 2012a, Jones et al 2012b, Marques et al 2016, Roesti et al 2015, Terekhanova et al 2014). These putative adaptive loci were dispersed on different chromosomes and some of them clustered as “islands of divergence”. A large portion (41%) of adaptive loci identified in global marine and freshwater stickleback comparisons were located in non-coding regions, while a small portion (17%) of them were found in coding regions (Jones et al 2012b). This indicates that changes in regulatory regions play a primary role in the adaptation of sticklebacks, which is consistent with the results from analyses of individual adaptive loci described above. In addition, chromosomal inversions may promote adaptation of sticklebacks as adaptive loci identified in marine-freshwater and lake-stream stickleback comparisons clustered on several genomic inversions (Jones et al 2012b, Roesti et al 2015). Lastly, a genomic survey of global marine and freshwater sticklebacks across the Northern Hemisphere demonstrated standing genetic variants carried by marine sticklebacks were repeatedly selected in the genomes of freshwater sticklebacks during adaptation, indicating the prominent role of standing genetic variation in stickleback adaptation (Jones et al 2012b).

1.6 Benthic and limnetic sticklebacks

1.6.1 Morphological divergence of benthic and limnetic sticklebacks

A special species pair of sticklebacks provides an exceptional model to study adaptation. While most of the rivers and lakes contain a single population of sticklebacks, species pairs evolved in at least five lakes in British Columbia, Canada (**Fig. 1.6a**) (Rundle & Schluter 2004). The limnetic ecotype (hereafter limnetics) usually lives in an open-water environment during the non-breeding season, while the benthic ecotype (hereafter benthics) lives in the littoral zone and never exploits open-water environments (McPhail 1984, McPhail 1992, McPhail 1994). Benthics and limnetics from different lakes show parallel morphological and diet divergence (Schluter & McPhail 1992). Limnetics feed on plankton while benthics eat small invertebrates. In addition, these two ecotypes are different in several morphological traits including body size, lateral plate number, gill raker number, gill raker length, gape width, and number of neuromasts (**Fig. 1.6b**) (McPhail 1994, Schluter & McPhail 1992, Wark & Peichel 2010). To adapt to the open-water environment and planktonic diet, limnetics have small and slim bodies, high armor plate counts, complete pelvic and dorsal spines, numerous long gill rakers, and small jaws. In contrast, benthics have large bodies, reduced armor plates, no armor spines, few and short gill rakes, and large jaws (McPhail 1992, Schluter & McPhail 1992). The divergence of morphological traits between benthics and limnetics has a strong genetic basis and can be retained in common lab settings (Hatfield 1997). It has been shown that the divergence was a result of competition for resources between two ecotypes in sympatry (Schluter 1994, Schluter & McPhail 1992). Therefore, the parallel morphological divergence between benthic and limnetic ecotypes provides strong evidence for the role of natural selection in their speciation and adaptation.

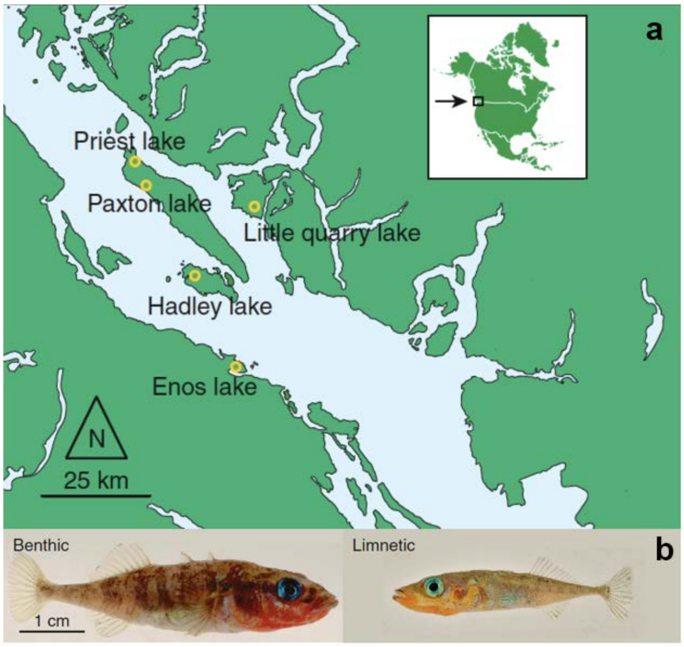

Figure. 1.6 | Geographic distribution and morphology of benthics and limnetics. **a,** The geographic locations of five lakes where benthics and limnetics are found together. **b,** The morphology of benthic and limnetic sticklebacks. Figure from (Roesti & Salzburger 2014).

1.6.2 Benthic and limnetic stickleback speciation

Strong reproductive isolation was found between sympatric benthics and limnetics repeatedly in different lakes (Rundle et al 2000). However, reproductive isolation was absent between the same ecotypes of different lakes. Furthermore, reproductive isolation between different ecotypes of the same lake is slightly higher than the isolation between different ecotypes from different lakes. This suggests that disruptive natural selection played a critical role in generating the reproductive isolation between these two ecotypes.

Evidence for both prezygotic and postzygotic isolation between benthics and limnetics have been documented. First, it has been found that body size and male nuptial color contributed to premating isolation between benthics and limnetics (Boughman et al 2005). Females of both ecotypes prefer to mate with conspecific males that have similar body sizes as themselves. The preference is stronger in benthic than limnetic females. In addition, limnetic females distinguish males by their nuptial coloration (Boughman et al 2005). The body color of sticklebacks helps them to be cryptic in their habitat, but male sticklebacks gain nuptial coloration during the breeding season (Boughman 2001). As limnetic sticklebacks breed in an environment where the water is clear, limnetic males display nuptial coloration of red throats, iridescent blue eyes, and blue or green backs. In contrast, benthic males develop nuptial coloration with dark black bodies because they breed in a darker environment (Boughman 2001). Limnetic females prefer to mate with males with brighter nuptial colors, which are found in conspecific males. Therefore, premating isolation by female preference in benthics is primarily determined by body size, while isolation in limnetics is decided by both body size and male nuptial coloration (Boughman et al 2005). The premating isolation between benthics and limnetics was repeatedly found in different lakes, suggesting that natural selection contributed to the formation of premating isolation between the ecotypes (Boughman et al 2005).

Postzygotic isolation has also been observed between benthics and limnetics. The divergence of morphological traits between benthics and limnetics substantially affect their survival in nature by allowing them to obtain food more efficiently in their own niche (Schluter 1993). Thus, the two ecotypes grow much faster in their respective environments and slower in the other’s (Schluter 1995). In contrast, hybrids of these two ecotypes have intermediate morphology and suffer the consequent reduction in feeding efficiency and growth rate in both the lab environment and the wild (Arnegard et al 2014, Hatfield & Schluter 1999, Schluter 1995). As the disadvantages of hybrids attribute to intermediate morphology but not intrinsic incompatibilities, this suggests there is postzygotic isolation between benthics and limnetics (Schluter 1993, Schluter 1995).

Advocates of sympatric speciation try to find evidence of it from sympatric species residing in isolated geographic areas, while advocates of allopatric speciation consider these species pairs as secondary contact of allopatric species after geographic changes (Coyne & Orr 2004). Thus, sympatric benthic and limnetic sticklebacks inhabiting in multiple isolated lakes is a good system to study the prevalence of sympatric vs. allopatric speciation.

Two hypotheses have been proposed to explain the evolutionary history of benthic and limnetic sticklebacks (Rundle & Schluter 2004). Because of the well-documented evidence for both premating and postmating isolation between benthics and limnetics, it is proposed that these two ecotypes evolved in sympatry within each lake, and people sometimes use these species-pairs as an example of sympatric speciation (Coyne & Orr 2004, Rundle & Schluter 2004). In contrast, McPhail proposed a double-invasion scenario that marine sticklebacks invaded the lakes on two separate occasions (McPhail 1993, Schluter & McPhail 1992). The first invaders evolved to be benthic specialists while the second invaders specialized in the limnetic habitat. It has previously been estimated that the second invasion occurred 1,500~2,000 years after the first one (Schluter & McPhail 1992).

Scientists have used genetic and genomic approaches to investigate the genetic relationship between benthic and limnetic sticklebacks. Two studies of the evolutionary history of benthics and limnetics using six microsatellite markers and a SNP genotyping array supported the double-invasion hypothesis (Jones et al 2012a, Taylor & McPhail 2000). Both studies identified features consistent with the predictions of the double-invasion hypothesis: polyphyletic origin of species-pairs in the same lake, lower heterozygosity of benthics than limnetics, and closer relationship of limnetics with marine sticklebacks than benthics.

1.7 Reverse speciation of Enos Lake benthics and limnetics

Enos Lake on Vancouver island, British Columbia, Canada is one of the five lakes (Paxton, Priest, Little Quarry, Enos, Hadley Lake) in which sympatric benthics and limnetics reside (Roesti & Salzburger 2014). McPhail (1984) first identified the sympatric stickleback ecotype pair in Enos Lake and showed ecotype pair in the lake has similar morphological divergence as benthic and limnetic ecotype pair in other lakes (McPhail 1984). In 2001, researchers found 12% of the sampled sticklebacks in Enos Lake have intermediate morphology, which should be classified as hybrids (Kraak et al 2001). Thus, they hypothesized that benthics and limnetics might have “collapsed” into a single hybrid swarm (reverse speciation). The study of sticklebacks in Enos Lake collected from 1977 to 2002 using morphological and genetic data showed the reverse speciation might start between 1994 and 1997 (Taylor et al 2006). It was hypothesized that the reverse speciation of Enos Lake benthics and limnetics was due to the introduction of crayfish (*Pascifasticus lenisculus*) to Enos Lake in the early 1990s, which might have destructed aquatic vegetation and reduced water clarity (Taylor et al 2006). A genetic study using microsatellite markers determined that the species “collapse” is due to the introgression from benthics to limnetics (Gow et al 2006), making the hybrid in Enos Lake was phenotypically similar to benthics and was able to consume foods of both benthics and limnetics (Rudman & Schluter 2016).

To preserve the species pairs, an effort was made by Dolph Schluter from 1988 to 1989. Enos Lake limnetics were introduced to the Murdo Frazer Pond in Murdo-Frazer Park in Vancouver, Canada. Sticklebacks were collected from the pond in 1997 and preserved in the lab and are used to represent Enos Limnetics in this thesis. In contrast, Enos Benthics sampled from Enos Lake itself in 2008 and preserved in ethanol are used in this thesis to represent Enos Benthic ecotypes.

1.8 Summary of my studies

Scientists have conducted intensive morphological and ecological studies on the speciation and adaptation processes of benthics and limnetics

(McPhail 1984, Rundle et al 2000, Rundle & Schluter 2004, Schluter & McPhail 1992). Furthermore, comprehensive quantitative trait locus (QTL) mapping of several important traits has been performed using the hybrids of benthics and limnetics (Arnegard et al 2014, Conte et al 2015). These ecological and genetic studies of benthics and limnetics have greatly improved our understanding of their speciation and adaptation. However, there are still several important aspects of the speciation and adaptation of benthics and limnetics which remain unknown. Firstly, as the model of speciation (sympatric vs. allopatric) of benthics and limnetics is subject to controversy, it is important to investigate their evolutionary history in more detail. In previous studies, the evolutionary history of the species pair was only inferred using genetic variations of mitochondrial DNA (mtDNA), microsatellite sites, and few thousand SNPs generated from a SNP genotyping array (Jones et al 2012a, Rundle & Schluter 2004). Secondly, parallel speciation, in which similar traits and reproductive isolation evolve in separate closely-related populations independently, provides strong evidence for the role of natural selection in evolution (Conte et al 2012). Benthics and limnetics are one of the classical examples of parallel speciation (Rundle et al 2000, Schluter & Nagel 1995). However, a comprehensive survey of how many genetic regions are repeatedly used by different species pairs of benthics and limnetics has been limited to just one study done by QTL mapping (Conte et al 2015). Thirdly, the genomic pattern of genetic divergence between benthics and limnetics is largely unknown. The genomic study of marine and freshwater sticklebacks revealed several islands of divergence in the genome (Jones et al 2012b). As islands of divergence are not universal in the genomes of related species (see **Section 1.4.2**), it is important to know whether benthics and limnetics also have islands of divergence and what evolutionary factors (i.e. selection, recombination, gene flow) shaped these islands. Fourthly, it is interesting to know if the sympatric species pairs used *de novo* mutations or standing genetic variation in their adaptation, Fifthly, as both benthics and limnetics live in freshwater lakes, it is interesting to investigate whether the sympatric ecotype pairs used the same set of adaptive loci as marine and freshwater sticklebacks. Sixthly, divergence in gene expression has been shown to have a critical role in both adaptation and speciation of several organisms (Stern & Orgogozo 2009, Wittkopp & Kalay 2012), especially the adaptation of freshwater sticklebacks (Jones et al 2012b). Nevertheless, the divergence of gene expression regulation remains to be determined in benthics and limnetics. Lastly, several traits have been identified to be important for the adaptation and speciation of benthics and limnetics (Arnegard et al 2014). However, knowledge of the genetic basis of these adaptive traits is still limited with only two genes regulating adaptive traits identified by QTL mapping (Chan et al 2010, Miller et al 2007). Although there have been efforts to comprehensively identify genomic regions controlling adaptive traits in benthics and limnetics using QTL mapping (Arnegard et al 2014, Conte et al 2015), these works suffered from low resolution of QTL mapping, which sometimes result in identifying regions too large to be informative (e.g., half a chromosome). In my dissertation, I set to resolve these questions using whole genome re-sequencing datasets of benthics and limnetics from four lakes (Paxton Lake, Priest Lake, Little Quarry Lake, Enos Lake) in British Columbia, Canada as well as RNA sequencing dataset of Paxton Lake benthics and limnetics.

In chapter 2, I study the genomic pattern of adaptive genetic variations in benthics and limnetics by analyzing whole genome re-sequencing data of six benthic and six limnetic individuals from each of the four lakes as well as 23 individuals each of Paxton Lake benthics and limnetics. I investigate the parallelism of genetic divergence between benthics and limnetics from different lakes. In addition, I identify regions with high genetic divergence and their distribution in the genomes of benthics and limnetics. Furthermore, I disentangle the factors that might contribute to the formation of genomic landscape of genetic divergence of benthics and limnetics. Finally, I detect genomic regions under positive selection in the genomes of benthics and limnetics, and compare the pattern of selection in these two species.

In chapter 3, I study the genetic basis of adaptation and speciation of benthics and limnetics. Firstly, I identify adaptive loci in benthics and limnetics and disentangled whether the adaptive variations of benthics and limnetics derived from *de novo* mutations or standing genetic variation, and whether benthics and limnetics used the same set of adaptive loci as marine and freshwater sticklebacks. Secondly, I analyze the biological functions of the adaptive loci in benthics and limnetics. Finally, I collaborate with a lab mate to dissect the function of two candidate adaptive regions in benthics and limnetics using enhancer essays.

In chapter 4, I study the evolutionary history of benthics and limnetics using whole-genome resequencing data. First, I identify the genetic relationship between benthics and limnetics in the context of marine and freshwater sticklebacks (210 individuals) and attempt to identify and characterize the populations sharing most ancestry of benthics and limnetics. Second, I identify the best-fit demographic model of Paxton Lake benthics and limnetics using simulation and historical effective population size (N_e) inference.

In chapter 5, I dissect the genomic pattern of *cis*-regulatory divergence in lab-created F1 hybrids of Paxton Lake benthics and limnetics. I study the functions of *cis*-regulatory genes that 1) show divergence between Paxton Lake benthics and limnetics and 2) are selected during adaptation of benthics and limnetics.

As stated previously, the reverse speciation of Enos Lake benthics and limnetics provides an excellent model to study the speciation and maintenance of the divergence between two species. In chapter 6, I determine the extent and genomic pattern of reverse speciation of Enos Lake benthics and limnetics. I compare the genetic divergence of the Enos Lake species pair to benthics and limnetics from other lakes as well as global marine and freshwater sticklebacks.

2 GENOMIC PATTERNS OF ADAPTIVE GENETIC VARIATION IN BENTHIC AND LIMNETIC STICKLEBACKs

2.1 Background and Aims

Identifying and analyzing adaptive loci in various organisms provides insight into how natural selection shapes the genome and individual traits during evolution (Wolf & Ellegren 2017). In addition, describing how adaptive loci are arranged in the genome is an important subject of evolutionary study and has become one of the most active areas in adaptation research (Faria et al 2014). For example, identifying the extent of linkage disequilibrium among adaptive loci can provide insight into how genomic architecture facilitates or constrains rapid adaptation (Barrett & Hoekstra 2011).

The understanding of the genetic basis of parallel morphological divergence in benthics and limnetics is still lacking. A study comparing QTLs controlling several important traits for benthic and limnetic adaptation showed that nearly half of the QTLs were reused during adaptation (Conte et al 2015). However, genes/loci contributing to similar traits may be identified as a single QTL, resulting in large QTLs that span up to half a chromosome (Savolainen et al 2013). Therefore, studying parallel genetic divergence of benthics and limnetics using QTL mapping may underestimate the number of parallel divergent regions. As described in **Section 1.2.3**, the genomic regions that are highly diverged between populations living in contrasting environments are considered to have been subject to positive selection (Vitti et al 2013). Using genomic approaches, highly divergent regions between populations can be identified in high resolution (Savolainen et al 2013). Therefore, it is important to comprehensively study the parallel genetic divergence of benthics and limnetics using genomic approaches.

Uncovering the genomic landscape of adaptive divergence in the genomes of closely related species is one of the central goals in adaptation research (Faria et al 2014). As described in **Section 1.4.2**, studies of closely related populations from one species that have adapted to divergent environments have identified a heterogeneous genomic landscape of genetic divergence with highly differentiated regions dispersed on a background of low divergence (Nosil et al 2009a). “Islands of genetic divergence”, which is extended regions with elevated divergence can be found in several bot not all species. The genomic study of marine and freshwater stickleback ecotypes across the Northern Hemisphere found several “islands of genetic divergence”, suggesting these “islands” are important for stickleback adaptation to freshwater environment (Jones et al 2012b). However, the genomic landscape of divergence between benthics and limnetics is largely unknown. Therefore, it is important to investigate the genomic landscape of adaptive divergence in benthics and limnetics.

Genomic regions of high divergence between closely related populations can be derived from selection, sorting ancestral alleles, or genetic drift (Nosil et al 2009a). Only divergent regions derived from divergent natural selection contribute to adaptation of populations. Thus, it is critical to identify genomic regions that are selected during benthics and limnetics adaptation. Nevertheless, only one study detected signals of selection in the genomes of benthics and limnetics using few thousand SNPs generated by SNP genotype array (Jones et al 2012a). As a result, it is important to identify and compare signals of selection in these two species using SNPs identified by whole genome resequencing.

In this chapter,

- I characterize the genomic composition of variation in benthics and limnetics by comparing site frequency spectrums and evaluating the divergence between them.
- I investigate the genetic basis of parallel morphological divergence between benthics and limnetics and identified the genomic landscape of divergence between these two species.
- I investigate the strength and type of selection as well as the origins of selective alleles in benthics and limnetics.

2.2 Sequencing and data generation

To investigate the adaptation of benthics and limnetics, six wild-caught benthics and six wild-caught limnetics from each of the four lakes (Paxton, Priest, Little Quarry, and Enos Lake) were whole-genome resequenced to an average coverage of 13.47x (**Appendix Table 1**). To increase the statistical power of several analyses in this thesis, 17 additional Paxton Lake benthics and limnetics were whole-genome sequenced to an average coverage of 26.66x (**Appendix Table 2**). In addition, six marine and six freshwater sticklebacks from Little Campbell River, Canada and River Tyne, Scotland were whole-genome sequenced, as Little Campbell River is geographically closed to these four lakes and samples from River Tyne were used in a previous genomic analysis of a global set of marine and freshwater sticklebacks (Jones et al 2012b). The average sequencing coverage was 17.41x for the Little Campbell River samples and 8.08x for the River Tyne samples (**Appendix Table 3**). Finally, to study the evolutionary history of benthics and limnetics, 186 individuals from a global set of marine and freshwater stickleback populations were whole-genome sequenced to an average coverage of 6.04x (**Appendix Table 4**).

All resequencing reads were aligned against the stickleback reference sequence assembly (gasAcu1) (Jones et al 2012b). After stringent filtering, high-quality SNPs were identified between the reference sequence and the resequenced individuals (see **Materials and Methods** for detail). Three SNP datasets were generated for the analyses in this thesis:

1. SNP dataset of benthics and limnetics from different lakes. Six benthic and six limnetics from each of the four lakes were included in this dataset. Moreover, six marine and six freshwater sticklebacks from Little Campbell River and River Tyne were included as reference. A total of 12,684,692 high-quality SNPs were identified between the reference sequence and the 72 individuals.
2. SNP dataset of Paxton Lake benthics and limnetics. Twenty-three Paxton Lake benthics and 23 Paxton Lake limnetics as well as 6 marine and 6 freshwater ecotypes from Little Campbell River and River Tyne were included in this dataset. In total, 10,655,570 high-quality

SNPs were identified between the reference sequence and the 70 individuals.

3. SNP dataset of benthics, limnetics, and global marine/freshwater sticklebacks. Six benthic and six limnetic individuals from each of the four lakes as well as 210 marine and freshwater stickleback individuals sampled across the Northern Hemisphere (including samples from Little Campbell River and River Tyne) were included in this dataset. A total of 21,175,919 high-quality SNPs were identified between the reference sequence and the 258 individuals.

2.3 Adaptive variations of benthics and limnetics

2.3.1 Evaluation of genomic composition of benthics and limnetics

Sympatric benthics and limnetics can interbreed and about 1% of stickleback individuals collected in the wild are possible hybrids between benthics and limnetics (Schluter & McPhail 1992). To ensure that the samples of benthics and limnetics from the different lakes were not hybrids, I first evaluated the genomic composition of benthics and limnetics from different lakes using principal component analysis (PCA) of genome-wide SNP data (**Fig. 2.1**). The first principal component (PC1) explains 11.78% of the variation in the genome and separates benthics and limnetics from all four lakes significantly (Paxton Lake: *P*-value = 1.62 x 10^{-14}, Priest Lake: *P*-value = 1.11 x 10^{-16}, Little Quarry Lake: *P*-value = 5.57 x 10^{-12}, Enos Lake: *P*-value = 2.09 x 10^{-9}, Tracy-Widom statistics). The second principal component (PC2) explains 8.1% of the variation in the genome and separates stickleback individuals by lakes. This suggests that the benthics and limnetics used in this study represent typical sympatric species pairs in the lakes, and can be used to study parallel benthic-limnetic speciation and adaptation. Interestingly, Enos Lake limnetics (ENSL) are shifted on PC1 towards Enos Lake benthics (ENSB), suggesting Enos Lake limnetics became more benthic-like in their genome, which might due to the increased gene flow between them. As described in **Section 1.7**, a group of Enos Lake limnetics was collected between 1988 and 1989 and transplanted to a small isolated pond for preservation. The samples of Enos Lake limnetics used in this study are individuals from the small pond, which are considered typical Enos Lake limnetics. The PCA reveals a closer genetic relationship between the benthics and limnetics from Enos Lake compared to the species pairs from the three other lakes. This suggests the increase of hybridization between benthics and limnetics started before 1988, which is earlier than the previous estimate of 1994 (Taylor et al 2006). Detailed analyses of reverse speciation of Enos Lake benthics and limnetics can be found in **chapter 6**.

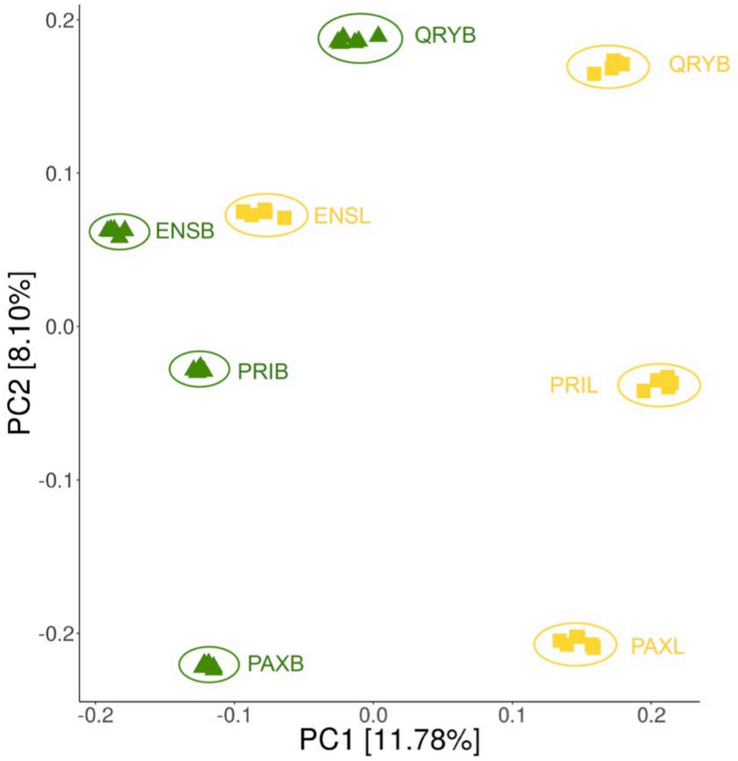

Figure 2.1 | Principal component analysis (PCA) of benthics and limnetics from different lakes. PCA was performed using genome-wide SNPs. The first principal component (PC1) separates benthics (green triangles) and limnetics (yellow squares) from different lakes. Enos Lake limnetics (ENSL) are shifted on PC1 towards Enos Lake benthics (ENSB), which is consistent with the gene flow from Enos Lake benthics to limnetics. PAXB: Paxton Lake benthics; PAXL: Paxton Lake limnetics; PRIB: Priest Lake benthics; PRIL: Priest Lake limnetics; QRYB: Little Quarry Lake benthics; QRYL: Little Quarry Lake limnetics; ENSB: Enos Lake benthics; ENSL: Enos Lake limnetics.

The genomic composition of 23 Paxton Lake benthics and 23 Paxton Lake limnetics was also evaluated with PCA using whole genome SNPs (**Fig. 2.2**). The first principal component (PC1) explains 28.01% of variation in the genome and separates benthic and limnetic sticklebacks significantly (*P*-

value < 1× 10^{-56}, Tracy-Widom statistics). Variation explained by the first and second principal components differ greatly, with the second principal component (PC2) only explaining 2.22% of the variation. Only limnetics separate on PC2, indicating that limnetic sticklebacks have higher genetic diversity than benthic sticklebacks.

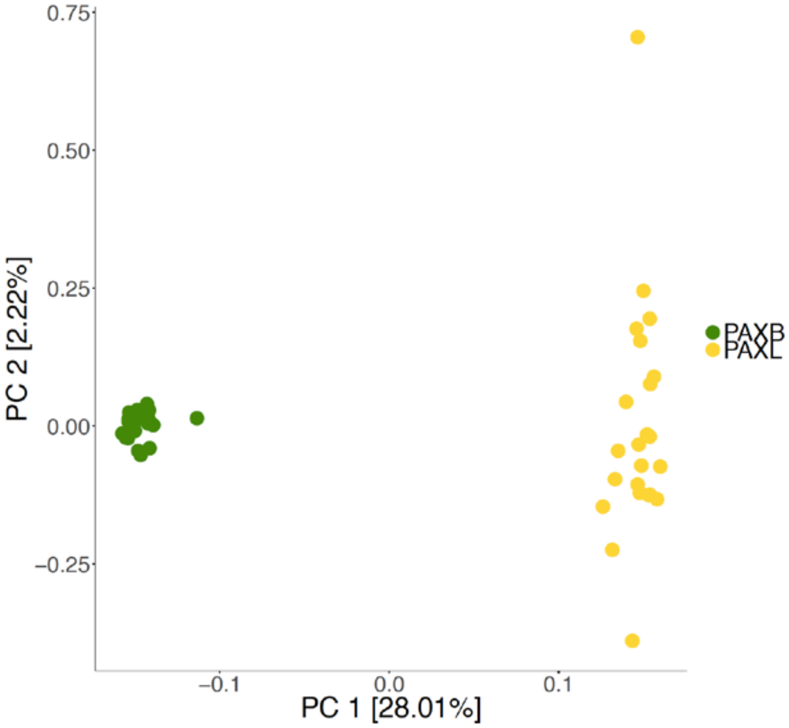

Figure 2.2 | Principle component analysis (PCA) of 23 Paxton Lake benthics (PAXB) and 23 Paxton Lake limnetics (PAXL). PCA was performed using genome-wide SNPs. The first principal component (PC1) separates Paxton Lake benthics and limnetics. The second principal component (PC2) separates different individuals of Paxton Lake limnetics.

2.3.2 Genomic variations of benthics and limnetics from different lakes

Genome-wide heterozygosity of benthics and limnetics was estimated using average heterozygosity (2pq) and nucleotide diversity (π). In addition, the number of variants observed only in a one individual of a population (singletons) was quantified in benthics and limnetics. A hybrid zone, which is a small geographic area where divergent populations encounter and hybridize, is an excellent system to study speciation because it provides empirical examples of divergence and gene flow (Hewitt 1988). Therefore, 5 hybrid zone marine and freshwater stickleback population pairs, which are populations from lower and upper reaches of the same river, were included in the analysis as reference (**Table 2.1**).

Table 2.1 Detailed information of hybrid zone marine and freshwater stickleback populations

Code	Population Name	Ecotype	Basin	Geographic Region	Country	Sample Size
LITC_DWN	Little Campbell River Downstream	Marine	Pacific	White Rock	Canada	6
LITC_UP	Little Campbell River Upstream	Freshwater	Pacific	White Rock	Canada	6
BIGR_DWN	Big River Downstream	Marine	Pacific	California	USA	5
BIGR_UP	Big River Upstream	Freshwater	Pacific	California	USA	5
BNMA	Bonsall Creek Downstream	Marine	Pacific	Vancouver Island	Canada	5
BNST	Bonsall Creek Upstream	Freshwater	Pacific	Vancouver Island	Canada	5
MIDF_DWN	Midfjardara River Downstream	Marine	Atlantic	Iceland	Iceland	5
MIDF_UP	Midfjardara River Upstream	Freshwater	Atlantic	Iceland	Iceland	5
TYNE_DWN	River Tyne Downstream	Marine	Atlantic	East Lothian	Scotland	6
TYNE_UP	River Tyne Upstream	Freshwater	Atlantic	East Lothian	Scotland	6

The mean heterozygosity (2pq) and π are higher in marine than in freshwater populations (2pq: 0.1731 versus 0.1405, π: 0.0022 versus 0.0019), and there are more singletons in the genomes of marine populations than freshwater populations (**Fig. 2.3**). A higher heterozygosity and more singletons in marine sticklebacks are consistent with a larger effective population size (N_e) in marine populations, as genetic drift cannot effectively remove or fix genetic variants in large populations (Hedrick 2005). Interestingly, freshwater ecotypes from Bonsall Creek (BNST) have a similar level of heterozygosity but fewer singletons than marine ecotypes (BNMA). This might result from gene flow between from marine to freshwater ecotypes in the river.

The mean heterozygosity (2pq) and π are higher in limnetics than in benthics (2pq: 0.1803 versus 0.1551, π: 0.0027 versus 0.0022), and limnetics have more singletons in their genomes than benthics (**Fig. 2.3**). This suggests limnetics have a larger N_e than benthics. There are fewer singletons in the genomes of Enos Lake limnetics than benthics. This might arise from the homogenizing effect of gene flow from benthics to limnetics during the process of reverse speciation in Enos Lake.

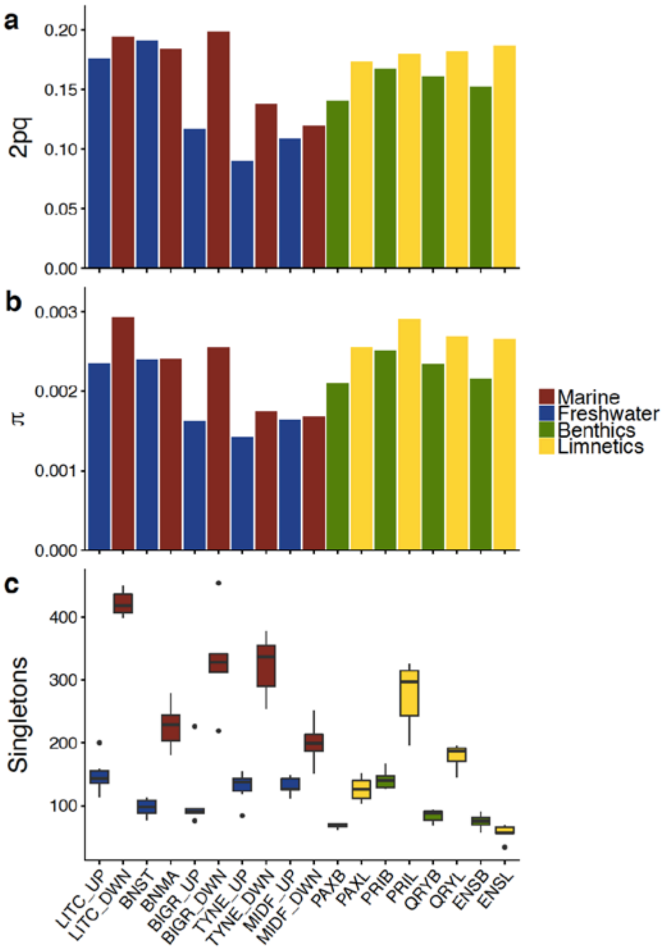

Figure 2.3 | Genome-wide genetic variation of benthics, limnetics, marine and freshwater populations. a, Average heterozygosity (2pq). **b,** Nucleotide diversity (π). **c,** Number of singletons per genome. Refer **Table 2.1** for population codes of marine and freshwater stickleback populations. PAXB: Paxton Lake benthics; PAXL: Paxton Lake limnetics; PRIB: Priest Lake benthics; PRIL: Priest Lake limnetics; QRYB: Little Quarry Lake benthics; QRYL: Little Quarry Lake limnetics; ENSB: Enos Lake benthics; ENSL: Enos Lake limnetics.

The pattern of linkage disequilibrium (LD) can be used to estimate recent N_e of a population as LD between pairs of SNPs depends on N_e and recombination rate at the same time. LD between variants further apart from each other reflects low recent N_e, as recombination cannot break down the linkage between SNPs effectively with a small population size (Tenesa et al. 2007). As natural selection can extend LD at target regions (see **Section 1.2.2**), I measured the LD between SNPs on putative "neutral" chromosome (chromosome XV) of benthics and limnetics as well as marine (LITC_DWN) and freshwater populations (LITC_UP) from Little Campbell River, Canada (**Fig. 2.4**). Chromosome XV is considered putatively "neutral" because there are no QTLs controlling adaptive traits of benthics and limnetics identified on this chromosome (Arnegard et al 2014, Conte et al 2015), and there are no divergent genomic regions between global marine and freshwater sticklebacks identified on this chromosome (Jones et al 2012b). LD decays with short physical distance (<20kb) in all studied populations. LITC_DWN has the shortest LD blocks, indicating that it has a larger N_e than other populations. Benthics and LITC_UP have longer LD blocks than limnetics and LITC_DWN population, which implies they have lower N_e than limnetics and LITC_DWN. LITC_UP has slightly shorter LD blocks than benthics, suggesting they have slightly higher recent N_e than benthics. Interestingly, Enos Lake limnetics have the longest LD blocks, indicating they experienced a more severe drop in N_e in recent years due to the reverse speciation event in the lake.

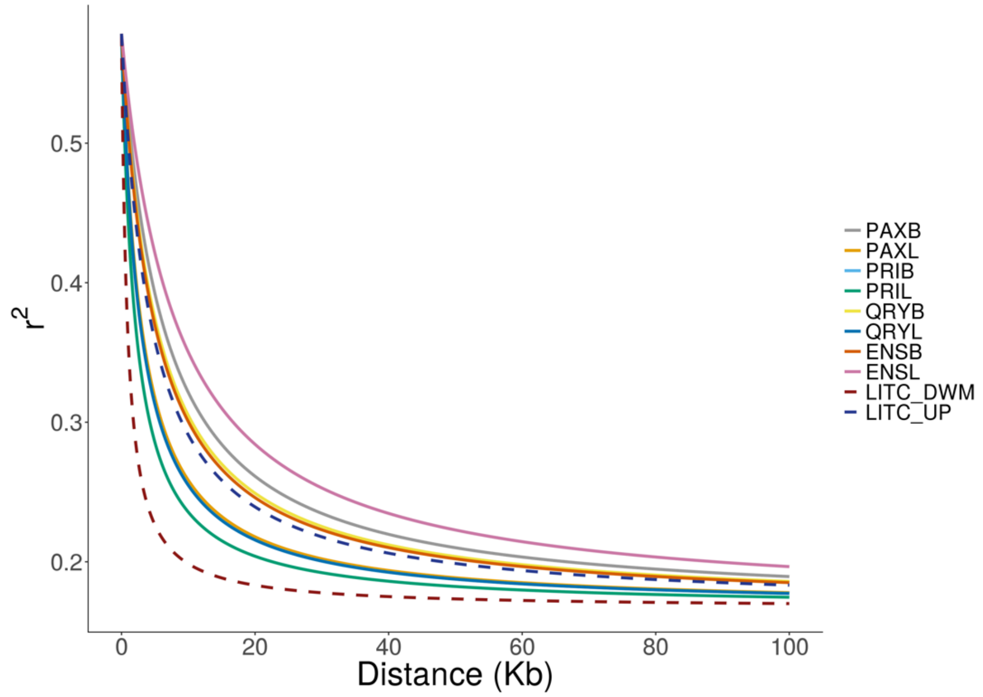

Figure 2.4 | Decay of Linkage disequilibrium (LD) on chromosome XV. LD was calculated and plotted for putative "neutral" chromosome (chromosome XV), which has no QTL mapped in benthics and limnetics from Paxton and Priest Lakes for several phenotypic traits (Arnegard et al 2014, Conte et al 2015).

Taken together, evaluating genomic variation and LD patterns of benthics, limnetics, as well as marine and freshwater sticklebacks found marine sticklebacks and limnetics had larger N_e than freshwater sticklebacks and benthics respectively. This suggests marine sticklebacks and limnetics have been through less of a population bottleneck than freshwater sticklebacks and benthics respectively. Marine sticklebacks having a larger N_e than freshwater sticklebacks is consistent with the current model of marine sticklebacks representing a large stable ancestral population from which freshwater sticklebacks have radiated in repeated small population bottlenecks (Bell & Foster 1994a).

2.3.3 Genomic divergence between benthics and limnetics from different lakes

Genome-wide genetic divergence (F_{ST}) between benthics and limnetics from different lakes ranges from 0.1388 to 0.23 (**Table. 2.2**), which is in the range of sympatric populations in the late stage of divergence (*Ficedula* flycatchers: F_{ST} = 0.291/0.303, *Heliconius* butterflies: F_{ST} = 0.287/0.292, Darwin's finches: F_{ST} = 0.23) but substantially higher than incipient sympatric populations (Lake Massoko African cichlid: F_{ST} = 0.038) (Burri et al 2015, Han et al 2017, Malinsky et al 2015, Nadeau et al 2012). The genetic divergence between benthics and limnetics is slightly higher than hybrid zone marine and freshwater stickleback populations (F_{ST} ranges from 0.048 to 0.204) This could have resulted from "higher rates" of gene flow between marine and freshwater sticklebacks compared to the benthics and limnetics, possibly due to the reinforcement of ecotype-specific mating preferences between the benthics and limnetics after they came into secondary contact.

Table 2.2 F_{ST} values of stickleback population pairs

Population pair	F_{ST}	Population pair	F_{ST}
LITC_UP vs. LITC_DWN	0.204	**PAXB vs. PAXL**	0.23
BNST vs. BNMA	0.137	**PRIB vs. PRIL**	0.21
BIGR_UP vs. BIGR_DWN	0.111	**QRYB vs. QRYL**	0.161
TYNE_UP vs. TYNE_DWN	0.106	**ENSB vs. ENSL**	0.139
MIDF_UP vs. MIDF_DWN	0.049		

Investigating the distribution of genome-wide genetic divergence (F_{ST}) can shed light on the degree of reproductive isolation and stage of speciation (Seehausen et al 2014). Therefore, I evaluated the distribution of genome-wide genetic divergence between benthics and limnetics from different lakes as well as hybrid zone marine and freshwater sticklebacks by calculating F_{ST} in 10kb non-overlapping windows. Most of the genomic regions have relatively low genetic divergence (F_{ST} < 0.2) between benthics and limnetics from different lakes, while a few genomic regions have high genetic divergence (F_{ST} > 0.5) (**Fig. 2.5**). The distributions of genetic divergence between hybrid zone marine and freshwater sticklebacks are similar to the distributions between benthics and limnetics. This distribution of genetic divergence is consistent with the late stage of speciation with gene flow (Martin et al 2013, Seehausen et al 2014). Therefore, both genome-wide mean and distribution of genetic divergence between benthics and limnetics as well as hybrid zone marine and freshwater sticklebacks suggest these two types of ecotype pairs are at the late stage of speciation with gene flow.

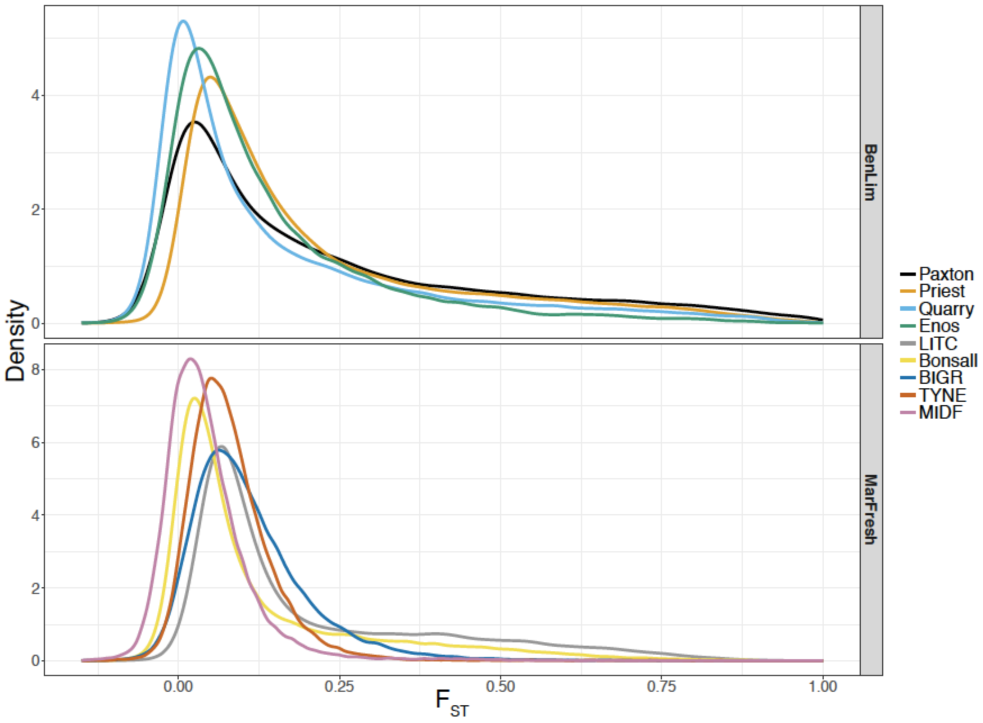

Figure 2.5 | Distribution of genetic divergence (F_{ST}) between benthics and limnetics (BenLim) from different lakes as well as hybrid zone marine and freshwater stickleback populations (MarFresh). F_{ST} was calculated in 10kb non-overlapping windows. LITC: Little Campbell River; Bonsall: Bonsall Creek; BIGR: Big River; TYNE: River Tyne; MIDF: Midfjardara River.

2.4 Parallel adaptive divergence between benthics and limnetics from different lakes

As described in **Section 1.4.2**, describing the number and distribution of adaptive loci is of fundamental importance and is one of the main subjects of evolutionary biology (Faria et al 2014). Empirical studies have demonstrated that adaptive phenotypic changes can be achieved by the modification of allele frequencies at a few loci of large effect, or at multiple loci of small to moderate effect (Lamichhaney et al 2015, van't Hof et al 2011). Therefore, to better understand the mechanism of a species' adaptation, it is important to disentangle the genetic architecture underlying phenotypic changes during adaptation. Genetic studies of repeated adaption of sticklebacks to diverse freshwater environments showed some of the important adaptive traits were regulated by one major locus with large effect size and several loci with small effect size (Colosimo et al 2004). In addition, Arnegard *et al.* (2014) investigated the genetic architecture of benthics and limnetics adaptation by mapping QTLs controlling several important adaptive traits and found most of the studied traits were regulated by several QTLs of moderate effect, suggesting the adaptation of benthics and limnetics has a polygenic basis (multiple loci involved in a single phenotypic changes) (Arnegard et al 2014). However, this study only used benthics and limnetics from one lake (Paxton Lake), and QTL mapping in sticklebacks has relatively limited powder due to their relatively small clutch sizes. Therefore, it is critical to investigate the genetic architecture of benthic and limnetic adaptation in fine scale using genomic approaches with the species pairs from multiple lakes.

2.4.1 Selection in benthics and limnetics from different lakes

Positive selection leaves a unique pattern of genetic variation in the genome. Amongst other things, it has the effect of increasing the frequency of advantageous alleles, resulting in an excess of high-frequency derived alleles within a population and strong genetic divergence between divergently adapted populations (Vitti et al 2013). Despite this, finding footprints of selection in the genome can be challenging when the number of loci responding to selection is large, the strength of selection relatively modest, and the substrate of selection is pre-existing genetic variation present at appreciable frequencies in the population (Stephan 2016). Such polygenic adaptation can leave subtle shifts in allele frequency at many loci across the genome (Stephan 2016). To explore the evidence for and the strength of selection in benthic and limnetic sticklebacks, I examined the genome wide F_{ST} relative to locus-specific differentiation and compared the shape of the site frequency spectrum.

To determine whether the high population divergence between stickleback populations evolved from natural selection or neutral demographic history, I evaluated the strength and prevalence of natural selection in stickleback populations by comparing genome-wide mean F_{ST} with extreme allele frequency differences in stickleback and compared this to human populations. The human genetic variant dataset (Phase 3) from the 1000 Genomes project (Altshuler et al 2015) was used for comparing genome-wide mean F_{ST} with extreme frequency differences in human. Fourteen human populations representing a wide geographic distribution and ancestry, and with a sample size equal to or greater than 6, were selected for the analysis (**Table 2.3**). To eliminate the effect of sample size variation between sticklebacks and human, 6 individuals were randomly selected in human populations with sample size greater than 6. Pairwise genome-wide F_{ST} and extreme allele frequency difference at individual loci were calculated for 14 human populations, benthics and limnetics from different lakes, and hybrid zone stickleback populations (**Fig. 2.6**). Long divergence time results in elevation of genome-wide genetic divergence, while strong positive selection increases the allele frequency difference at specific genomic loci (Vitti et al 2013). Thus, in two population pairs that have similar genome-wide mean F_{ST}, the population pair that has more genomic regions with extreme allele frequency difference underwent stronger divergent natural selection (Coop et al 2009). Almost all stickleback population pairs have more regions of the genome showing extreme allele frequency difference compared to human population pairs with similar mean F_{ST}. This is unlikely to be caused by neutral demographic processes such as population bottlenecks during divergence because these would have the effect of increasing the genome-wide F_{ST} as well as locus specific allele frequency differences (Coop et al 2009). It is therefore likely that stickleback populations have been subject to stronger divergent selection than human populations. Interestingly, the extreme allele frequency differences are larger in pairwise comparisons between species than within species of benthics and limnetics, which indicates benthics and limnetics evolved as a response to strong divergent natural selection.

Table 2.3. Detailed information of human populations used in the analysis of pair-wise mean F_{ST} and extreme frequency difference

No.	Population Description	Super Population	Sample number used in analysis
1	African Caribbean in Barbados	AFR	6
2	African Ancestry in SW USA	AFR	6
3	Luhya in Webuye, Kenya	AFR	6
4	Mende in Sierra Leone	AFR	6
5	Finnish in Finland	EUR	6
6	British from England and Scotland	EUR	6
7	Toscani in Italy	EUR	6
8	Chinese Dai in Xishuangbanna	EAS	6
9	Chinese in Beijing	EAS	6
10	Japanese in Tokyo	EAS	6
11	Bengali in Bangladesh	SAS	4
12	Gujarati Indians in Houston	SAS	6
13	Indian Telugu in the U.K	SAS	4
14	Kink in Ho Chi Minh City, Vietnam	SAS	6

Note: The human genetic variant dataset (Phase 3) generated by 1000 Genomes Project consortium was obtained from its website (Altshuler et al 2015). The human variant dataset was generated using whole-genome sequencing with a mean coverage of 7.4x. AFR: African population, EUR: European population, EAS: East Asian population, SAS: South Asian population

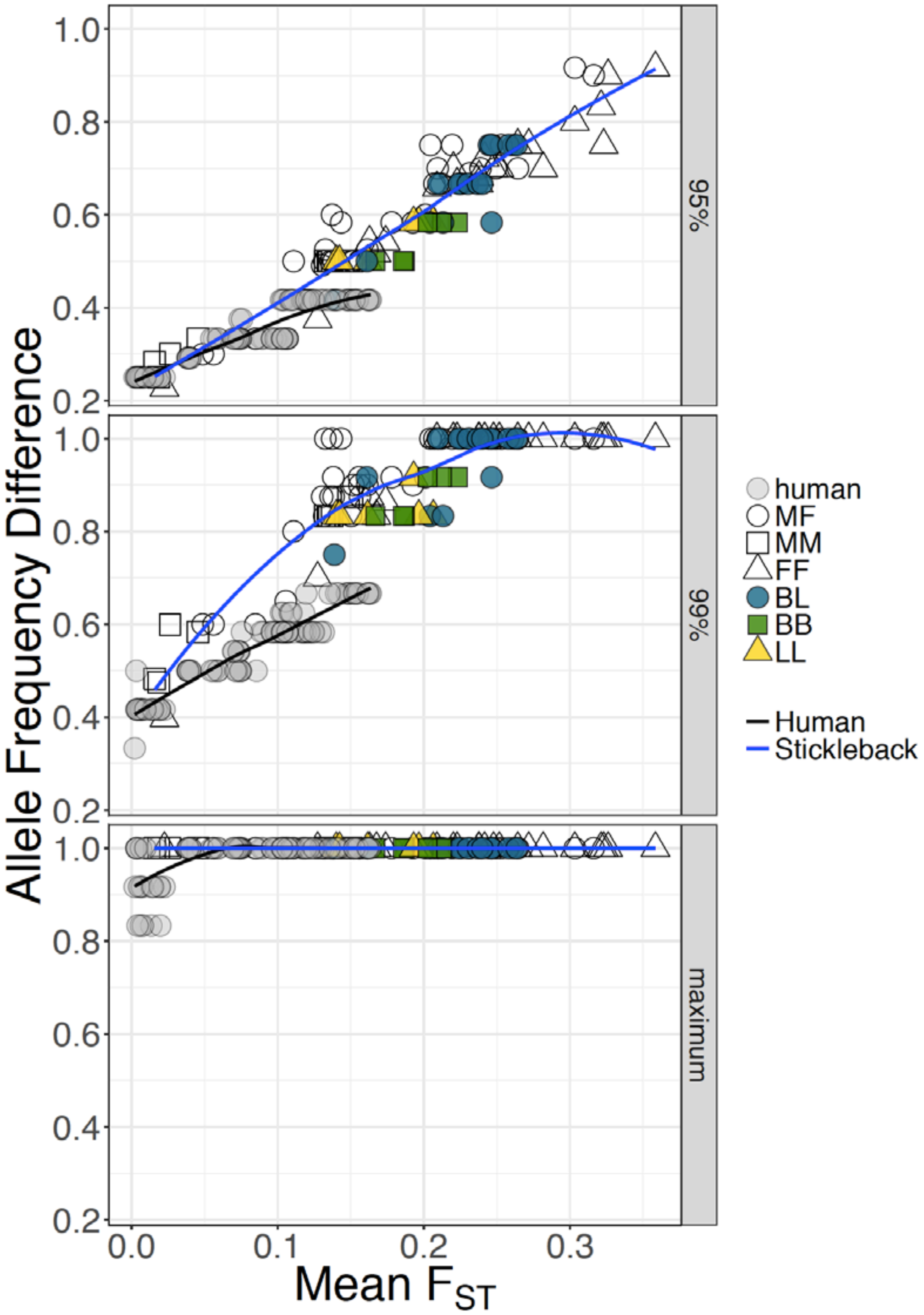

Figure 2.6 | The relationship of genome-wide mean F_{ST} and extreme allele frequency difference between populations of sticklebacks and human. Genome-wide mean F_{ST} is plotted on x-axis and extreme allele frequency difference is plotted on y-axis. The loess regression lines of sticklebacks and human are plotted in blue and black. Pairwise comparisons were performed for marine-freshwater (MF), marine-marine (MM), freshwater-freshwater (FF) ecotypes as well as benthics-limnetics (BL), benthics-benthics (BB), limnetics-limnetics (LL).

As described in **Section 1.2.1**, if a population experienced strong positive selection during evolution, the site frequency spectrum would shift to high-frequency alleles (Fay & Wu 2000). In contrast, negative selection removes deleterious mutations and prevents the mutations from reaching common frequency in the population, which leads to an excess of low-frequency alleles (Tajima 1989). To determine the types of selection that benthics and limnetics have been subject to during evolution, I calculated the unfolded site frequency spectra of benthics and limnetics from all four lakes (**Fig. 2.7a-d**). In addition, a joint (two-dimensional) site frequency spectrum was generated using 23 Paxton Lake benthics and 23 Paxton Lake limnetics for a better comparison (**Fig. 2.7e**). There are more high-frequency derived alleles in the genomes of benthics than limnetics, whereas limnetics have more low-frequency derived alleles than benthics. This suggests benthics have been subject to stronger positive selection, while limnetics experienced more negative selection during evolution.

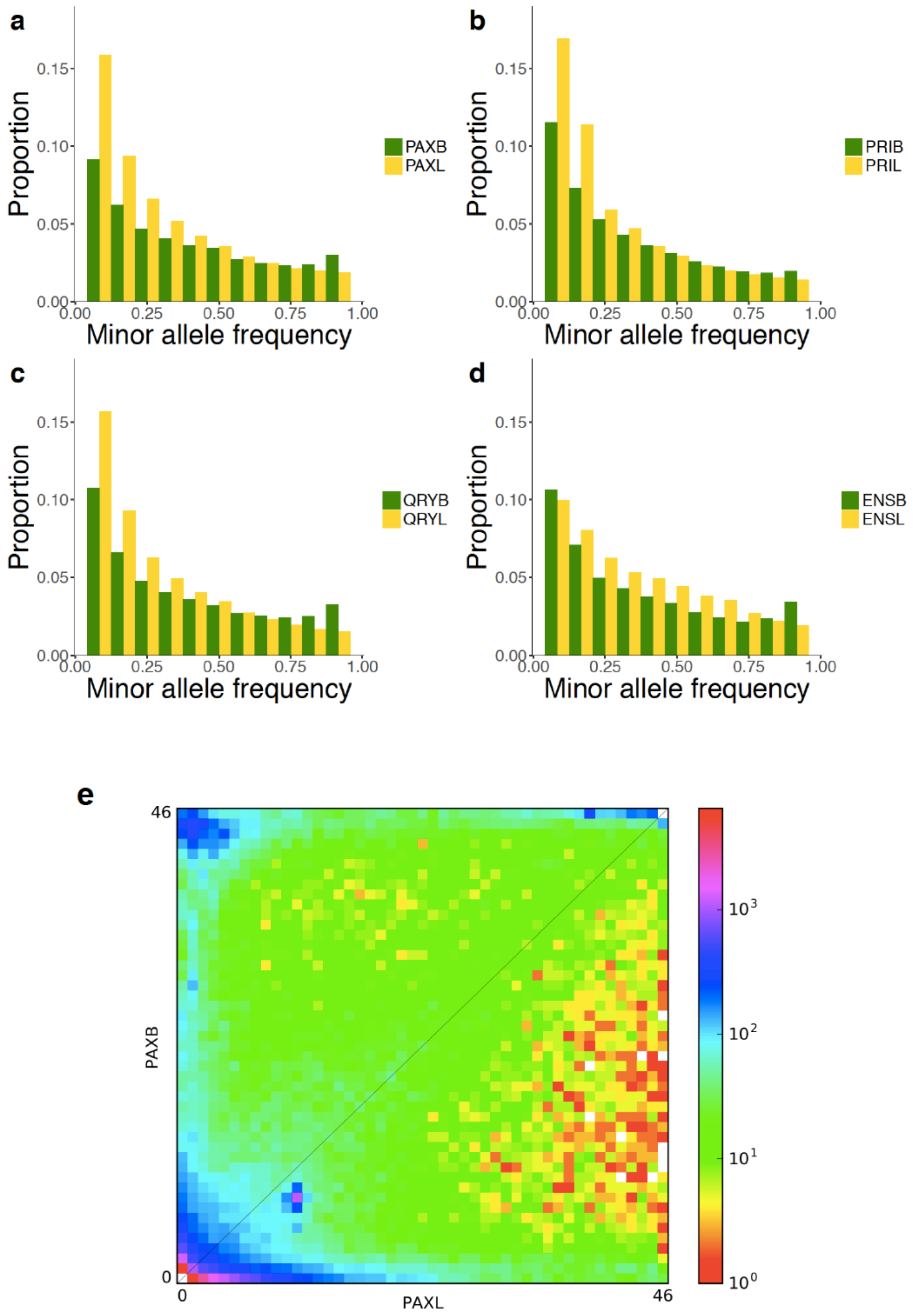

Figure 2.7 | Site frequency spectrum of benthics and limnetics from different lakes. Unfolded site frequency spectrums were calculated for 6 benthics and 6 limnetics from Paxton Lake (PAXB, PAXL) (**a**), Priest Lake (PRIB, PRIL) (**b**), Little Quarry Lake (QRYB, QRYL) (**c**), and Enos Lake (ENSB, ENSL) (**d**). **e**, Joint site frequency spectrum of 23 Paxton Lake benthics (PAXB) and 23 Paxton Lake limnetics (PAXL).

2.4.2 Pattern of parallel genomic divergence between benthics and limnetics from different lakes

Adaptation may occur via *de novo* mutation or by the reuse of pre-existing (“standing”) genetic variation (Messer & Petrov 2013). Previous studies suggest a large role for standing genetic variation in stickleback adaptation. For example, Jones et al (2012) showed that as much as 30% of loci underlying divergent adaptation of a given marine-freshwater ecotype pair is reused in parallel in independent marine-freshwater divergence events across the Northern Hemisphere (Jones et al 2012b). In addition, a previous QTL analysis estimated that 48.8% of QTLs controlling morphological divergence between benthics and limnetics from Paxton and Priest Lakes were shared in parallel, providing strong evidence for ecological adaptation (Conte et al 2015). As benthics and limnetics from different lakes showed parallel divergence for several morphological traits (Schluter & McPhail 1992), it is highly likely that benthics or limnetics from all four lakes used similar genetic variation during their adaptation to similar environments.

To investigate the parallel genomic divergence between benthics and limnetics from all four lakes, I first evaluated the genetic divergence between benthics and limnetics from all four lakes using the previously proposed cluster separation score (CSS) (Jones et al 2012b). CSS is a modified version of the widely used $F_{ST,}$ and measures the genetic divergence between populations by taking the genetic variation within populations into account. CSS scores were calculated by subtracting the mean of π between two individuals from different populations by the mean of π between two individuals from the same populations in sliding windows (size: 2,500bp; step: 500bp) across the chromosomes for each species pair from the four lakes.

CSS is highly correlated in the pairwise comparison of species pairs from the Paxton, Priest, and Little Quarry Lakes, but not Enos Lake (**Fig. 2.8**). CSS of benthics and limnetics from Paxton Lake and Priest Lake has the highest correlation (Spearman correlation: R = 0.66, *P*-value $< 1 \times 10^{-22}$) (**Fig. 2.8a**), and the correlation is lower for CSS of benthics and limnetics from Paxton Lake and Little Quarry Lake (Spearman correlation: R = 0.57, *P*-value $< 1 \times 10^{-22}$) (**Fig. 2.8b**). CSS of benthics and limnetics from Priest Lake and Little Quarry Lake has the lowest correlation (Spearman correlation: R = 0.54, *P*-value < 1 x 10^{-22}) (**Fig. 2.8c**). The correlation of CSS of benthics and limnetics from Enos Lake and each of the other three lakes is lower than the correlation between CSS of species pairs from each of these three lakes (Paxton vs. Enos: R = 0.43, *P*-value < 1 x 10^{-22}; Priest vs. Enos: R = 0.41, *P*-value < 1 x 10^{-22}, Quarry vs. Enos: R = 0.35, *P*-value < 1 x 10^{-22}) (**Fig. 2.9**). In addition, there are fewer genomic regions with elevated genetic divergence in benthics and limnetics from Enos Lake than from each of the other three lakes (**Fig. 2.9**).

Taken together, the pairwise comparisons of CSS showed that species pairs from different lakes had similar patterns of genetic divergence, indicating parallel morphological divergence has genetic basis. Benthics and limnetics from Enos Lake have fewer genomic regions with elevated divergence compared to the species pairs from each of the other three lakes. This might be due to the increased hybridization and gene flow between these two species.

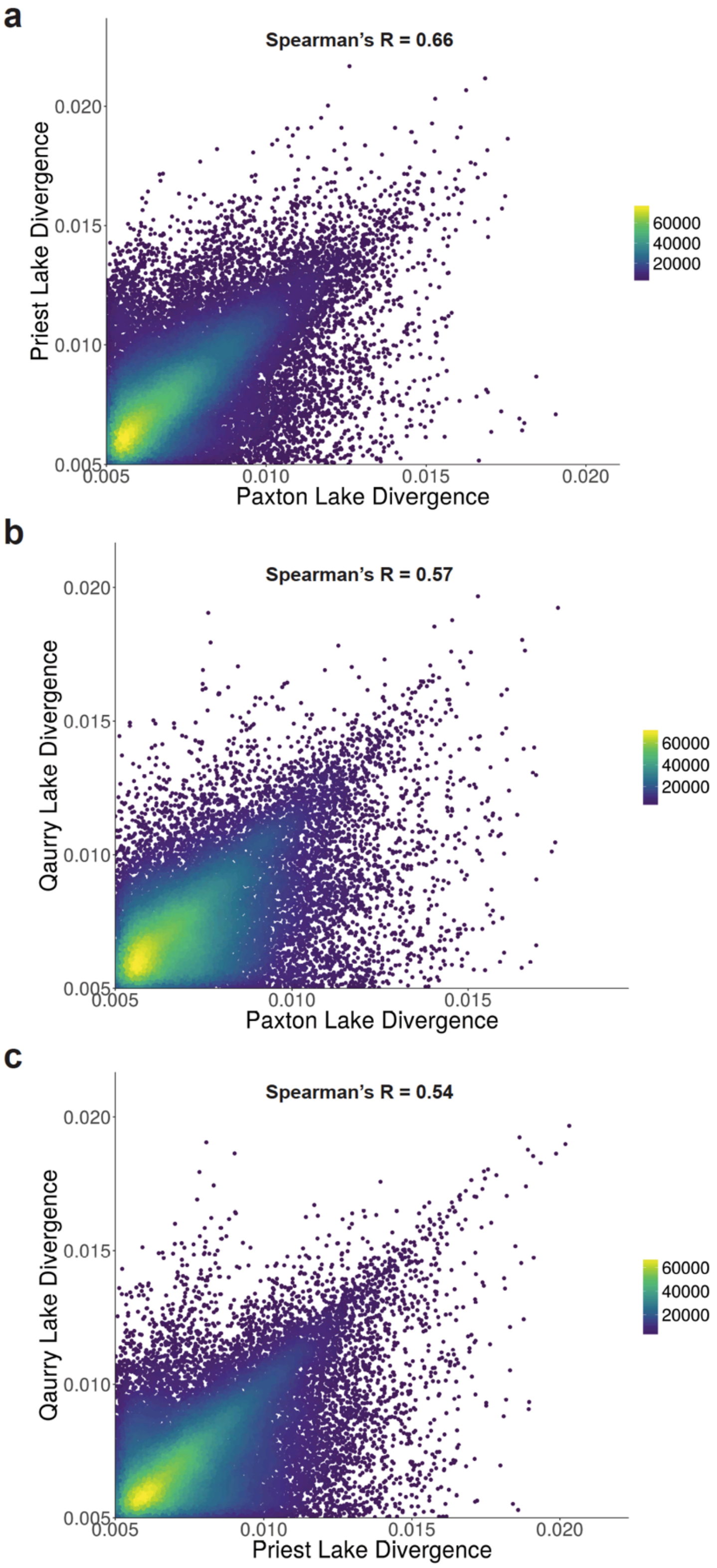

Figure 2.8 | Correlation of cluster separation score (CSS) in 10kb windows among species pairs from Paxton Lake, Priest Lake, and Little Quarry Lake. High correlations are found in each comparison (Spearman' correlation, Paxton vs. Priest: R = 0.66, *P*-value < 1 x 10^{-22}; Paxton vs. Quarry: R = 0.57, *P*-value < 1 x 10^{-22}; Priest vs. Quarry: R = 0.54, *P*-value < 1 x 10^{-22})

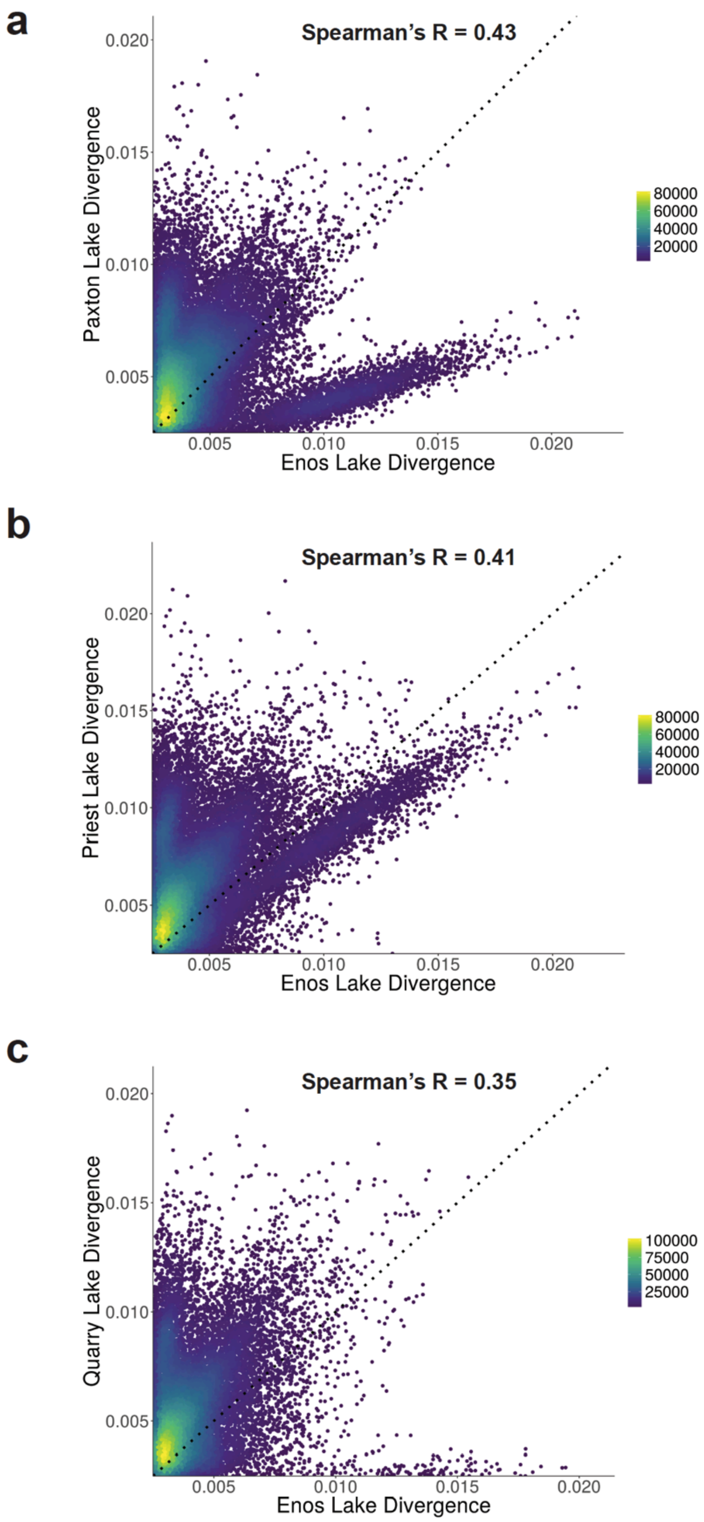

Figure 2.9 | Correlation of cluster separation score (CSS) in 10kb windows between species pairs from Enos lake and each of the other three lakes (Paxton, Priest, Little Quarry Lake). Relatively low correlations are found in each comparison (Spearman' correlation, Paxton vs. Enos: R = 0.43, *P*-value < 1 x 10^{-22}; Paxton vs. Quarry: R = 0.41, *P*-value < 1 x 10^{-22}; Priest vs. Quarry: R = 0.35, *P*-value < 1 x 10^{-22})

The genomic regions that are consistently highly diverged between two species contribute to their adaptation, as genetic drift is unlikely to fix the same alleles repeatedly (Elmer & Meyer 2011). Benthics and limnetics from different lakes have similar genomic patterns of divergence. Therefore, it is possible to identify genomic regions contributing to their adaptation by identifying regions that are highly diverged between these two species across different lakes. Although increased hybridization and gene flow homogenized several genomic regions in our samples of Enos Lake benthics and limnetics, one study still identified morphological divergence between individuals from these two species sampled until 1997 (Taylor et al 2006). As the samples of Enos Lake limnetics used in this study were derived from a population collected between 1988 and 1989 and preserved in a separate small pond, analyzing these samples should identify genomic regions contributing to their morphological divergence.

To identify the genomic landscape of divergence between benthics and limnetics, I evaluated genetic divergence between benthics and limnetics from all four lakes across the genome using CSS. CSS scores were calculated in 926,407 overlapping windows (2,500bp; step size: 500bp) across the chromosomes. Benthics or limnetics from all four lakes were combined as one population to identify parallel divergent regions between these two species. Numerous divergent genomic regions were identified between benthics and limnetics across all lakes (**Fig. 2.10**). Through large permutation testing (1 million times for each window), I identified 132,720 windows that are significantly diverged from the neutral expectation (empirical *P*-value = 0), indicating that 14.32% of the genome is diverged in parallel between benthics and limnetics. These overlapping windows correspond to 4,325 non-overlapping genomic regions, which I refer to as "parallel divergent regions". In addition, a total of 636,217 windows (68.7% of the genome) are not diverged from neutral expectation (empirical *P*-value > 0.05). These overlapping windows correspond to 9,063 non-overlapping genomic regions, which are considered as regions with no parallel divergence between benthics and limnetics ("parallel non-divergent regions").

The genomic regions that are diverged in parallel between benthics and limnetics from all four lakes show a non-random pattern of distribution: 1) some of the chromosomes have substantially more divergent regions than other chromosomes (**Appendix Table 5**); 2) divergent regions cluster and form 25 “islands of divergence” that are each larger than 250kb (median: 301,999bp; mean: 362,679bp; range: 252,499bp to 684,999bp) and are distributed over only six chromosomes (chrI, chrVII, chrIX, chrXVII, chrXVIII, chrXIX) as well as the pseudo-chromosome of unanchored scaffolds (chrUn) (**Fig. 2.10**, **Appendix Table 6**).

It has been proposed that “islands of genetic divergence” can be formed through “divergent hitchhiking” (see **Section 1.4.2**)(Nosil et al 2009a). On the other hand, large “islands of divergence” can also be formed if genetic hitchhiking occurs in genomic regions with low recombination rate (Nachman & Payseur 2012). The “islands of divergence” identified in benthics and limnetics from all four lakes are unlikely to arise from genetic hitchhiking as the neutral variants linked to the adaptive loci may not be shared among populations from different lakes. Gene flow can homogenize genetic variation in genomic regions that are not contributed to the adaptation, resulting in regions of low divergence. There are a few (1%) natural occurred hybrids between benthics and limnetics found in the wild (Schluter & McPhail 1992), indicating there is gene flow between these two species. Therefore, it is highly likely that the “islands of genetic divergence” identified in the cross-lake benthic and limnetic analysis have evolved from the interaction between natural selection and gene flow.

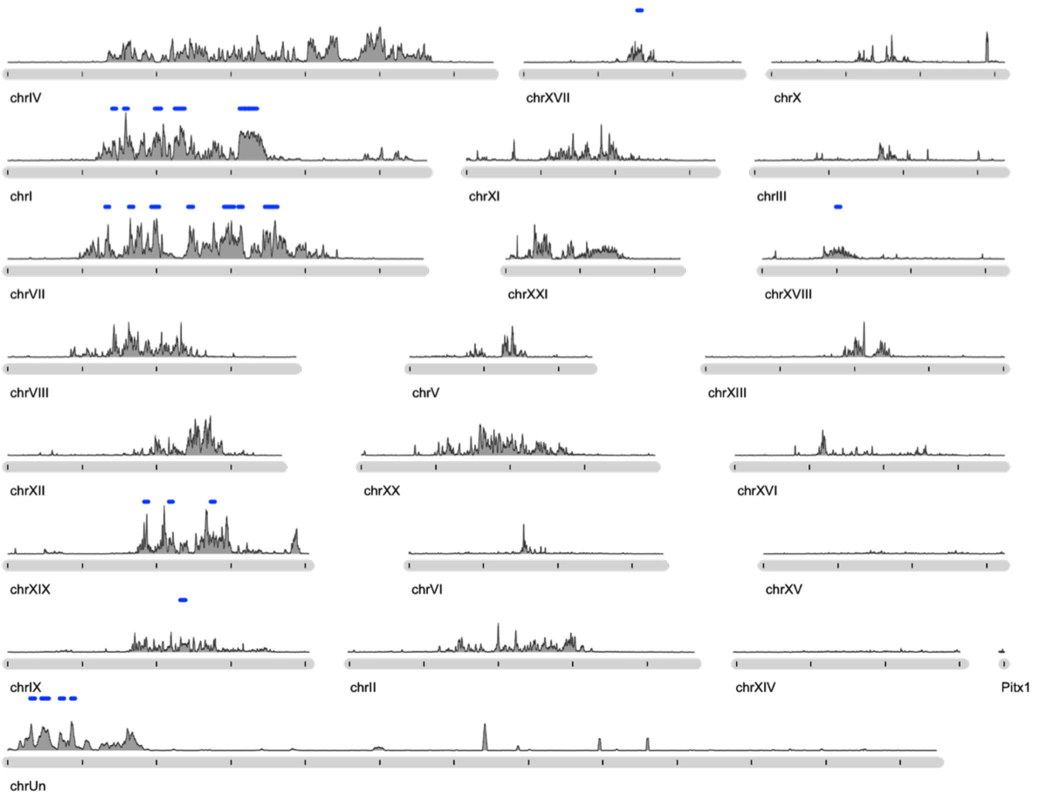

Figure 2.10 | Genomic pattern of divergence between benthics and limnetics from all four lakes. Genetic divergence was evaluated using cluster separation scores (CSS), which was calculated in 2,500bp overlapping windows with 500bp step size across all chromosomes. Each grey bar represents one chromosome and ticks in the bar indicates 5 Mb intervals. The blue bars on top of the chromosomes indicate "island of genetic divergence". As the genomic region containing *Pitx1* locus is not assembled in the reference sequence, this region is denoted as a separate "chromosome" in the graph.

Benthics and limnetics have diverged in their pelvic morphology. While limnetics have a pelvic spine, some benthics exhibit a reduction in their pelvic structures (McPhail 1994). The phenomenon of pelvic spine reduction exists only in benthics from Paxton Lake and Little Quarry Lake (McPhail 1994). Genetic study of pelvis spine reduction in sticklebacks demonstrated that the recurrent reduction of pelvic spine in diverse freshwater stickleback populations is due to independent *de novo* deletions in an enhancer (*Pel*) of the *Pitx1* locus (Chan et al 2010). In addition, deletion in this enhancer has been shown to contribute to the divergence of pelvic morphology between Paxton Lake benthics and limnetics (Chan et al 2010). As pelvic spine reduction does not exist in benthics across all four lakes, CSS scores of benthics and limnetics from different lakes are not high for the genomic region containing *Pitx1* locus in genome-wide distribution (**Fig. 2.10**). However, the windows containing *Pel* enhancer but not *Pitx1* locus show substantially higher CSS scores than other windows in the region (**Fig. 2.11**), indicating the *Pel* enhancer region is diverged between benthics and limnetics (even though not in all four lakes). As parallel divergence is a strong indicator of natural selection, the divergence between benthics and limnetics at the *Pel* enhancer region should result from selection. This is consistent with the result of a previous analysis suggesting that reduction of pelvic spine in freshwater sticklebacks (including benthics) is due to positive selection in the *Pel* enhancer region (Chan et al 2010).

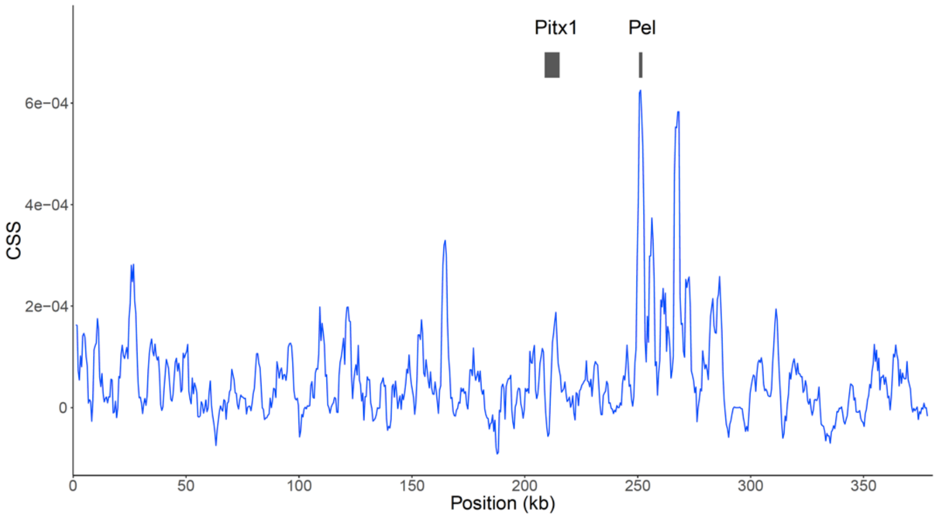

Figure 2.11 | Cluster separation scores (CSS) of benthics and limnetics from all four lake across *Pitx1* region. As the genomic region containing the *Pitx1* locus is not assembled in the stickleback reference sequence, improved sequences of bacterial artificial chromosomes (BACs) spanning the *Pitx1* locus were downloaded and concatenated for the analysis. CSS was calculated in 2,500bp overlapping windows with 500bp step. *Pitx1* locus and *Pel* enhancer are denoted as rectangles on top of the plot.

2.4.3 Identifying genomic regions under position selection in benthics and limnetics

The genomic regions that are consistently highly diverged between benthics and limnetics should contribute to their adaptation. However, high genetic divergence of genomic regions between populations can result from divergent selection in both populations or strong selection in only one of the populations. Therefore, it is important to identify genomic regions under positive selection in benthics and limnetics using methods based on other signatures of selection (i.e. allele frequency, linkage disequilibrium). As both benthics and limnetics cohabit in freshwater lakes, it is likely that some alleles or haplotypes that are important for general freshwater adaptation are selected in both species. Detecting signatures of selection in benthics and limnetics separately can identify regions where 1) divergent haplotypes were selected in benthics and limnetics (divergent selection) or 2) similar haplotypes were selected in these two species (directional selection). Thus, I identified genomic regions under positive selection in benthics and limnetics separately using methods based on allele frequency spectrum.

I used SweepFinder 2 to detect complete selective sweeps in benthics and limnetics using the composite likelihood ratio (CLR) statistic (DeGiorgio et al 2016) (see **Section 1.2.1** for detail description of CLR statistic). Benthics or limnetics from all four lakes were combined as one population in the analysis to identify genomic regions that consistently showed signatures of selective sweeps. Applying this approach to benthics and limnetics pooled across lakes involves testing for genomic regions where the pooled site frequency spectrum deviates from a neutral distribution. The null hypothesis in this approach states that the pooled site frequency spectrum follows a neutral model, which is reasonable because the site frequency spectrum under neutral expectation is only determined by mutation rate. However, this may be prone to false positives where population structure causes deviations in the site frequency spectrum. Regardless, the strongest CLR signatures will be achieved at regions of the genome where benthics and limnetics from all four lakes show signatures consistent with selection (excess of high frequency derived alleles)

CLRs were calculated in 2,500bp non-overlapping windows for the pooled benthics and limnetics from all four lakes respectively. Several genomic regions with extreme CLRs were identified in the pooled samples of benthics or limnetics from all four lakes, suggesting they were repeatedly selected during adaptation (**Fig. 2.12**). There are substantially more genomic regions with extreme CLRs in benthics from all four lakes (cross-lake benthics) than in limnetics from all four lakes (cross-lake limnetics). Interestingly, 1,410 out of 1,852 genomic regions (76.1%) with extreme CLRs in cross-lake benthics (top 1% in the genome-wide distribution) overlap with parallel divergent regions, while only 342 out of 1,852 genome regions (18.5%) with extreme CLRs in cross-lake limnetics (top 1% in the genome-wide distribution) overlap with parallel divergent regions. In addition, CLRs of cross-lake benthics at parallel divergent regions are significantly higher than parallel non-divergent regions ($P < 2.2\times10^{-16}$, two tailed Mann-Whitney U test), whereas CLRs of cross-lake limnetics at parallel divergent regions are significantly smaller than in parallel non-divergent regions ($P < 2.2\times10^{-16}$, two tailed Mann-Whitney U test).

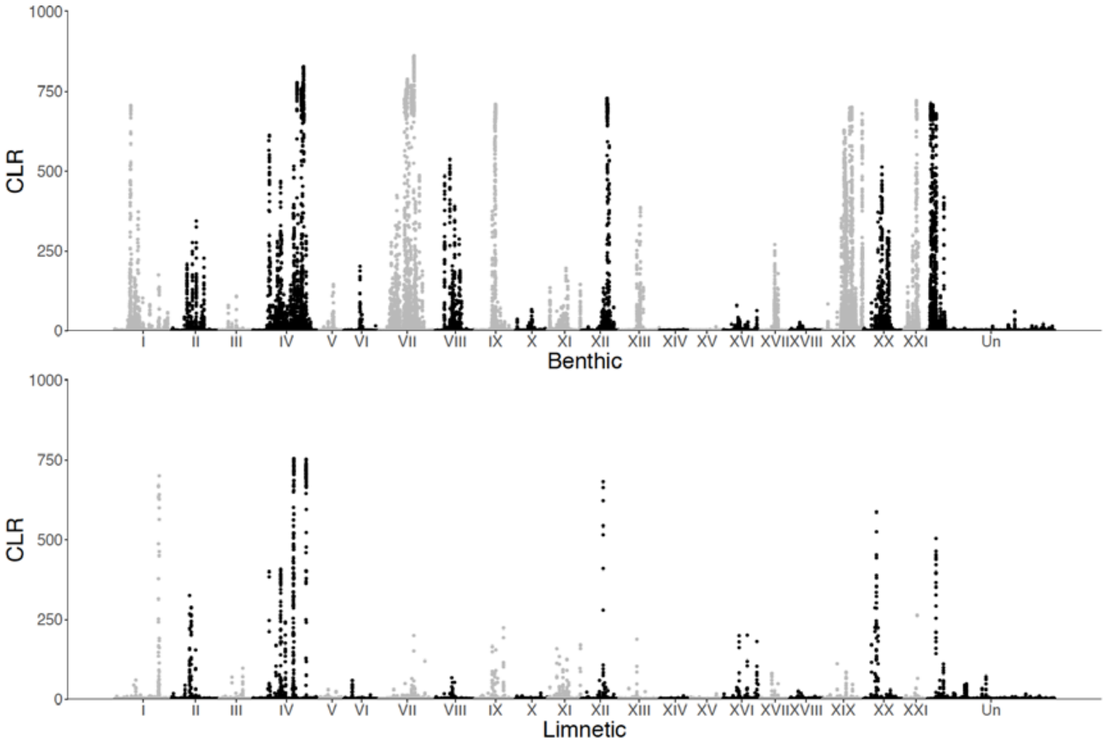

Figure 2.12 | Selective sweep in benthics or limnetics from all four lakes. Genomic regions identified as having been subject to a selective sweep based on their extreme composite likelihood ratio (CLR) (Kim & Stephan 2002). There are more regions under selective sweep in benthics than in limnetics.

CLR identifies selective sweep based on the significant deviation of site frequency spectrum from neutral expectation. As standing genetic variants segregate in the population for a long time, recombination can break down the linkage between these variants and other neutral variants. Therefore, the standing genetic variants can be carried by different haplotypes. The sweep of standing genetic variants can increase the frequency of multiple haplotypes in the population. As I used pooled samples of cross-lake benthics or limnetics in the analysis, limnetics from different lakes might carry different ancestral haplotypes if the ancestral alleles are selectively favored in limnetics at parallel divergent regions. Thus, CLRs of cross-lake limnetics might be low at parallel divergent genomic regions. To investigate whether derived or ancestral alleles are selected in benthics and limnetics at parallel divergent regions, I evaluated the genetic divergence between marine sticklebacks (marine stickleback ecotypes from Little Campbell River and River Tyne) and each of benthics and limnetics from all four lakes at parallel divergent regions. The genetic divergence was evaluated using F_{ST} in 2,500bp non-overlapping windows. Most of the windows have low divergence (F_{ST} < 0.2) between cross-lake limnetics and marine sticklebacks, while F_{ST} values of cross-lake benthics and marine sticklebacks range from small (F_{ST} <0.2) to large (F_{ST} > 0.5) (**Fig. 2.13**). This suggests the derived and ancestral alleles are selectively favored in cross-lake benthics and limnetics separately. The strong selection of derived haplotypes in benthics from different lakes contributes to the divergence between benthics and limnetics.

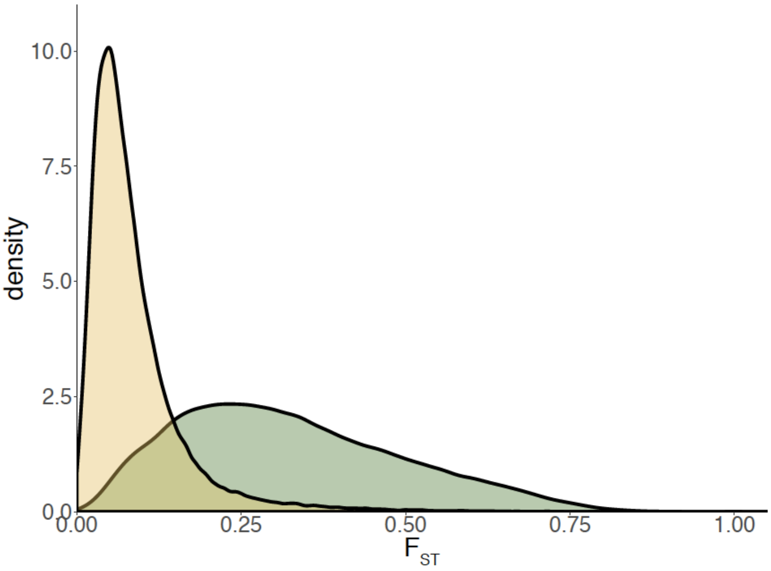

Figure 2.13 | Distribution of genetic divergence between marine sticklebacks and each of cross-lake benthics (green) and limnetics (yellow) at parallel divergent regions. Genetic divergence was evaluated using F_{ST} in 2,500bp non-overlapping windows. Most of the windows have low divergence ($F_{ST} < 0.2$) between cross-lake limnetics and marine sticklebacks, indicating limnetics are carrying ancestral alleles at these regions. Genetic divergence between cross-lake benthics and marine sticklebacks ranges from low ($F_{ST} < 0.2$) to high ($F_{ST} > 0.5$), suggesting benthics carry ancestral and derived alleles at these regions.

As the CLRs are more powerful in detecting selective sweeps on derived alleles (Pennings & Hermisson 2006), extreme CLRs in both cross-lake benthics and limnetics might indicate that strong selection of derived alleles occurred in both species. In total, 100 out of 1,852 genomic regions have extreme CLRs in both cross-lake benthics and limnetics. Most of these regions (benthics: 65.6%, limnetics: 54.3%) have high divergence ($F_{ST} > 0.5$) between marine sticklebacks (marine stickleback ecotypes from Little Campbell River and River Tyne) and cross-lake benthics or cross-lake limnetics, indicating derived haplotypes are selected in both benthics and limnetics (**Fig. 2.14a**). In addition, the genetic divergence between cross-lake benthics and limnetics is low ($F_{ST} < 0.2$) at most of these regions (84.6%)(**Fig. 2.14b**). Therefore, similar derived haplotypes were selected in both benthics and limnetics at the regions with extreme CLRs in both species, indicating these derived haplotypes are important for both benthic and limnetic adaptation. As both benthics and limnetics live in freshwater environments, these haplotypes might contribute to adaptation to the freshwater environment.

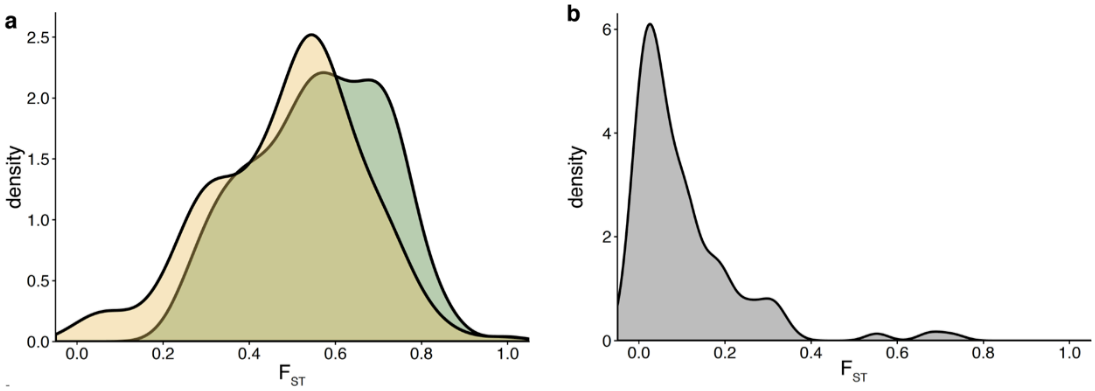

Figure 2.14 | Distribution of genetic divergence at genomic regions with extreme CLRs in both cross-lake benthics and cross-lake limnetics. Genetic divergence was evaluated using F_{ST} in 2,500bp non-overlapping windows. **a,** genetic divergence between marine sticklebacks and each of cross-lake benthics (green) and limnetics (yellow) at genomic regions with extreme CLRs in both species pairs. Most of the regions have high divergence ($F_{ST} > 0.5$) between marine sticklebacks and each of cross-lake benthics and limnetics. **b,** genetic divergence between cross-lake benthics and limnetics at genomic regions with extreme CLRs in both species pairs. Most of the windows have low divergence ($F_{ST} < 0.2$) between benthics and limnetics.

2.5 Adaptive divergence between Paxton Lake benthics and limnetics

Previous QTL studies of benthics and limnetics from Paxton and Priest Lake showed that 40% of QTLs regulate phenotypic divergence in one lake but not the other (Conte et al 2015), suggesting there are some uniquely divergent genomic regions between benthics and limnetics from each lake. In addition, pairwise comparisons of CSS scores of benthics and limnetics from different lakes showed there were several genomic regions that are diverged between benthics and limnetics from one of the lakes (**Fig. 2.8 and 2.9**), indicating there are unique patterns of genomic divergence between benthics and limnetics from individual lakes. Therefore, it is important to investigate the pattern of genetic divergence of the species pair from a single lake.

2.5.1 Pattern of genetic divergence between Paxton Lake benthics and limnetics

To determine the pattern of genetic divergence between benthics and limnetics from an individual lake, I evaluated the genetic divergence between 23 Paxton Lake benthics and 23 Paxton Lake limnetics using CSS. CSS was calculated in 926,407 overlapping windows (2500 bp, step size: 500bp). Numerous divergent regions were identified in the genome of Paxton Lake benthics and limnetics (**Fig. 2.15**). Surprisingly, the divergence between Paxton Lake benthics and limnetics at 481,577 windows is significantly deviated from neutral expectation (empirical *P*-value = 0, permutation test), indicating more than half of the genome (51.98%) is diverged between Paxton Lake benthics and limnetics. On the other hand, only 236,111 windows (25.5% of the genome) are not diverged from between Paxton Lake benthics and limnetics (empirical *P*-value > 0.05, permutation test).

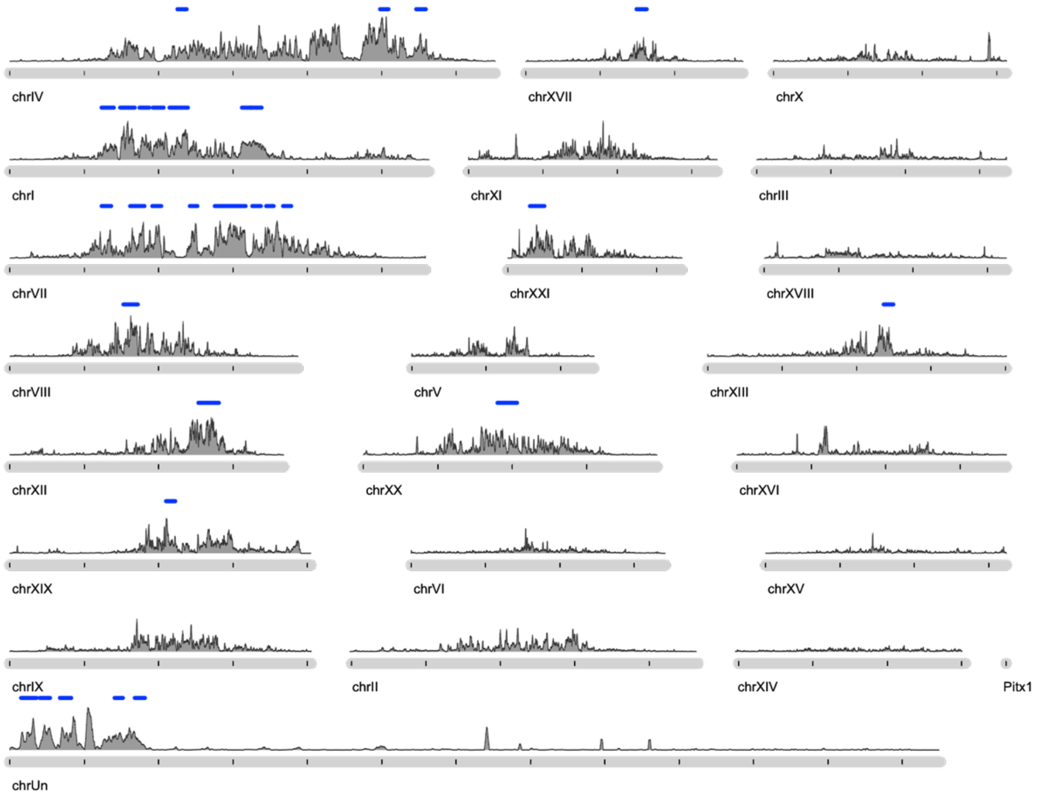

Figure 2.15 | Genomic pattern of genetic divergence between Paxton Lake benthics and limnetics. Cluster separation scores (CSS) were calculated in 2,500bp overlapping windows with 500bp step size across all chromosomes. Each grey bar represents one chromosome and ticks in the bar indicates 5 Mb intervals. The blue bars on top of the chromosomes indicate “islands of genetic divergence”.

Similar to the genomic pattern of parallel genetic divergence between benthics and limnetics from all four lakes, the divergent genomic regions of Paxton Lake benthics and limnetics cluster into 32 “islands of divergence” that span more than 500kb on several chromosomes, with four of them spanning more than 1Mb (Mean: 752,968bp; Median: 649,499bp; range: 500,999bp to 1,509,999bp) (**Appendix Table 7**).

2.5.2 Selection in Paxton Lake benthics and limnetics

Different methods of selection detection have power to identify signatures of selection that occur at different times in history. In addition, these methods have varying power to detect selection on *de novo* mutation or standing genetic variants. To compile a comprehensive landscape of selection, I detected selection in Paxton Lake benthics and limnetics using two different approaches based on different signatures of selection.

2.5.2.1 Detecting selection based on site frequency spectrum using sweepFinder2

To identify genomic regions under positive selection in Paxton Lake benthics and limnetics, I first calculated CLRs for 23 Paxton Lake benthics and 23 Paxton Lake limnetics in 2,500bp non-overlapping windows using sweepFinder2 separately. Unlike selective sweep detection in benthics and limnetics from all four lakes using SweepFinder2, both Paxton Lake benthics and limnetics have several regions with extreme CLRs in the genome, indicating these regions were selected in Paxton Lake benthics or limnetics (**Fig. 2.16**).

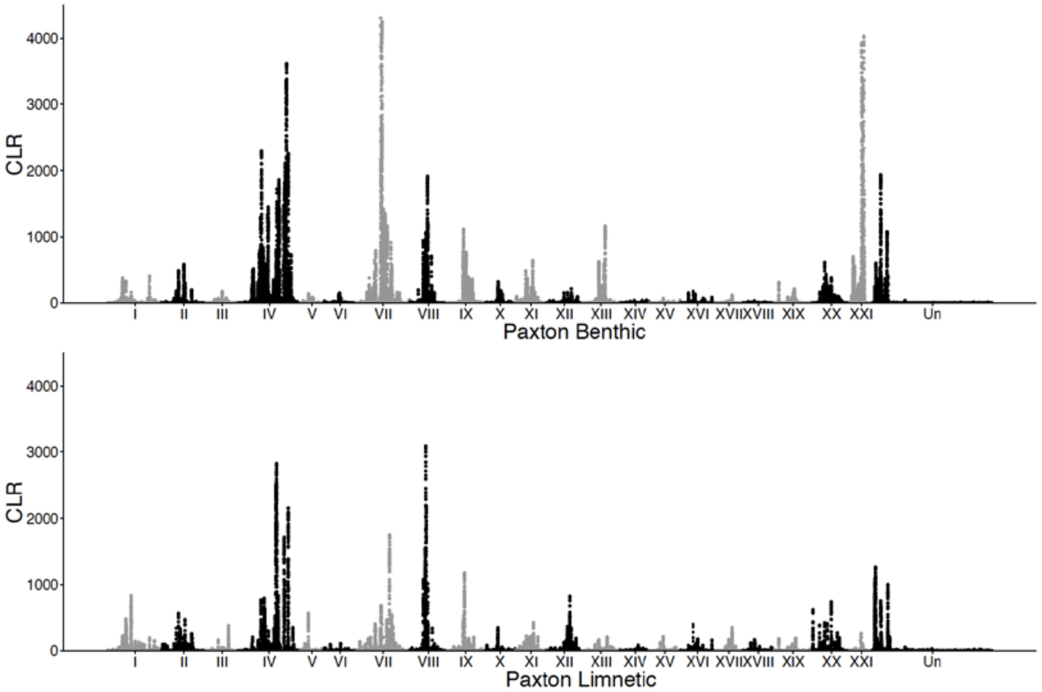

Figure 2.16 | Selective sweep in Paxton Lake benthics and limnetics. Selective sweep were detected using composite likelihood ratio (CLR) along chromosomes. Large CLR scores indicate strong signals of selection.

To investigate the contribution of natural selection to the divergence between Paxton Lake benthics and limnetics, I looked for overlapping genomic windows (2,500bp, step size: 500bp) having extreme CSS (top 0.5% in genomic distribution) between Paxton Lake benthics and limnetics as well as extreme CLR scores (top 0.5% in genomic distribution) in each of Paxton Lake benthics and limnetics. There are more genomic windows having both extreme CSS and CLR scores in Paxton Lake benthics (486 windows) than limnetics (290 windows)(**Fig. 2.17**). These windows cover 384,443bp (0.0083% of the genome; 57 genomic regions) and 247,955bp (0.054% of the genome, 45 genomic regions) of the genomes of Paxton Lake benthics and limnetics respectively (**Appendix Table 8 and 9**). This indicates that the divergence between Paxton Lake benthics and limnetics resulted from selective sweeps in both species, but predominantly resulted from sweep in Paxton Lake benthics.

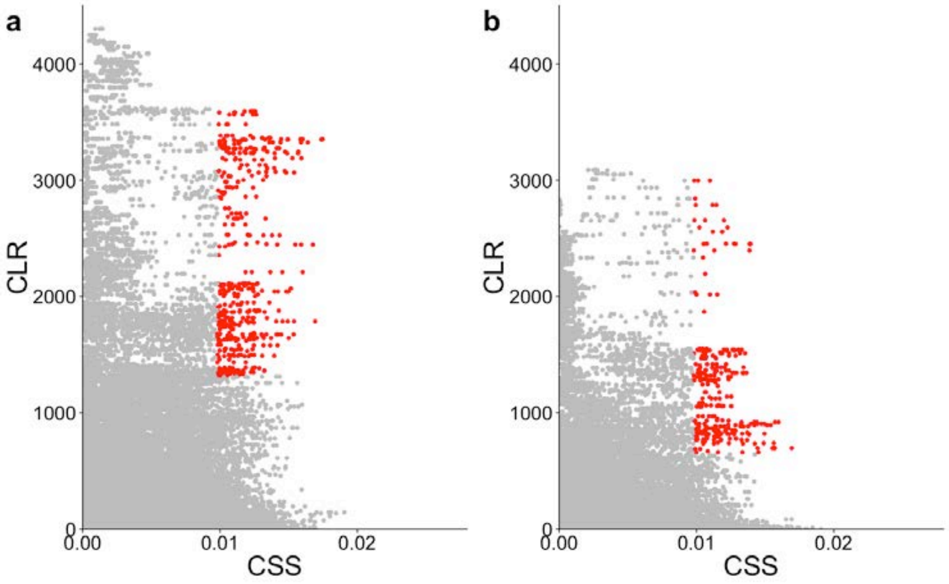

Figure 2.17 | Comparison of CSS and CLR scores in Paxton Lake benthics and limnetics. a, Paxton Lake benthics, highly divergent regions (top 0.5%, CSS>0.0098) with extreme CLR score (top 0.5%, CLR>1,315) are highlighted in red. **b,** Paxton Lake limnetics, highly divergent regions (top 0.5%, CSS>0.0098) with extreme CLR score (top 0.5%, CLR>647) are highlighted in red.

There are 78 genomic regions with extreme CLRs (top 0.5% in genomic distribution) in both Paxton Lake benthics and limnetics. Similar to the analysis in benthics and limnetics from all four lakes, most of these regions (Paxton Lake benthics: 89.7%, Paxton Lake limnetics: 75.6%) have large divergence (F_{ST}) between marine sticklebacks (marine stickleback populations from Little Campbell River and River Tyne) and each of the Paxton Lake benthics and limnetics (**Fig. 2.18a**). The majority of these regions (70.9%) have low divergence ($F_{ST} < 0.2$) between Paxton Lake benthics and limnetics (**Fig. 2.18b**). This suggests similar derived haplotypes were selected in both Paxton Lake benthics and limnetics at these regions. Interestingly, 10 genomic regions (12.7%) have large divergence between Paxton Lake benthics and limnetics, indicating divergent derived haplotypes were selected in these two species. Genes or functional elements in these regions may play an important role in the adaptation of Paxton Lake benthics and limnetics to their own environmental niches. Detailed analysis of the regions where divergent derived haplotypes are selected in Paxton Lake benthics and limnetics can be found later in **Section 3.2.2**.

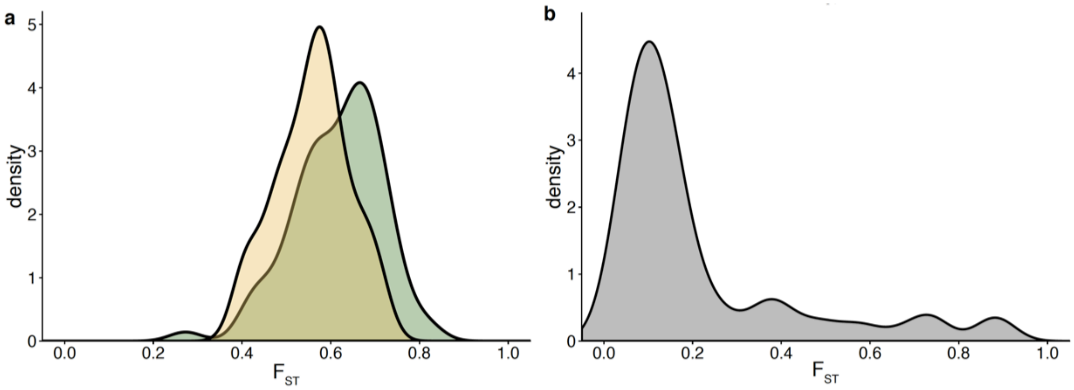

Figure 2.18 | Distribution of genetic divergence at genomic regions with extreme CLRs in Paxton Lake benthics and limnetics. Genetic divergence was evaluated using F_{ST} in 2,500 bp non-overlapping windows. **a,** genetic divergence between marine sticklebacks and each of Paxton Lake benthics (green) and limnetics (yellow) at genomic regions with extreme CLRs in both species pairs. Most of the regions have high divergence (F_{ST} > 0.5) between marine sticklebacks and each of Paxton Lake benthics and limnetics. **b,** genetic divergence between Paxton Lake benthics and limnetics at genomic regions with extreme CLRs in both species pairs. Most of the windows have low divergence (F_{ST} < 0.2) between benthics and limnetics.

2.5.2.2 Detecting selection based on linkage disequilibrium (LD) using nS_L statistic

I also identified genomic regions that underwent selection in Paxton Lake benthics and limnetics using the nS_L statistic. Integrated haplotype score (iHS) compares extensions of haplotypes carrying ancestral and derived core alleles (see **Section 1.2.2**)(Voight et al 2006). Differences between the extension of haplotypes carrying derived or ancestral alleles indicate selection on either *de novo* mutations or standing genetic variation. The calculation of iHS requires a genetic map to eliminate the effect of variation in recombination rate across chromosomes. The nS_L statistic uses the same approach as iHS to detect selection but measures the length of haplotype homozygosity between a pair of haplotypes in terms of the number of variations in other haplotypes in the genomic region, which can define the boundaries of haplotypes more accurately than inferring local recombination rate from a recombination map (Ferrer-Admetlla et al 2014). Large positive nS_L scores indicate selection on ancestral alleles and large negative nS_L scores indicate selection on derived alleles.

nS_L scores were calculated for SNPs with a minor allele frequency > 5% in 23 individuals each of Paxton Lake benthics and limnetics. The ancestral alleles of SNPs were determined according to the major genotype of marine ecotypes from Little Campbell River and River Tyne. Simulations have shown that it is more powerful to detect selective sweeps in windows that contain several SNPs with significant nS_L scores (Voight et al 2006). In addition, a sweep on derived/ancestral alleles sometimes increases the frequency of linked ancestral/derived alleles (Voight et al 2006). Therefore, I take the absolute values for nS_L scores of Paxton Lake benthics and limnetics and calculated the mean of absolute nS_L scores in 926,509 overlapping window (2500bp, step size: 500bp) (**Fig. 2.19**). A large mean absolute nS_L suggests a strong signal of positive selection. There are more genomic regions with large mean absolute nS_L in Paxton Lake benthics than in limnetics, indicating positive selection is more prevalent in Paxton Lake benthics than in limnetics.

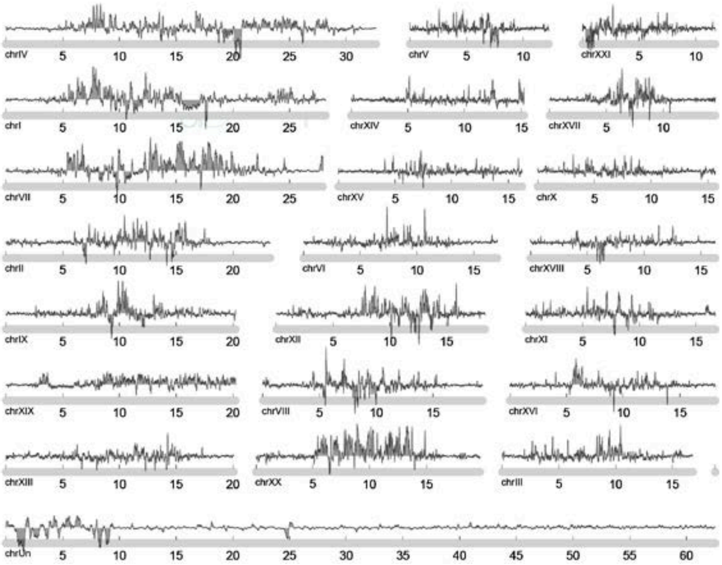

Figure 2.19 | Window mean of absolute nS_L in Paxton Lake benthics and limnetics. nS_L scores were calculated for SNPs with minor allele frequency > 5%. The mean of absolute nS_L scores of Paxton Lake benthics and limnetics in overlapping windows (2,500bp, step: 500bp) were calculated. The positive value indicates higher mean absolute nS_L in Paxton Lake benthics, while negative value indicates higher mean absolute nS_L in Paxton Lake limnetics.

To identify SNPs that were under selective sweep, I performed a large permutation analysis to calculate the empirical *P*-value for each test SNP in Paxton Lake benthics and limnetics. In total, 24,061 and 6,397 SNPs were identified as under positive selection in Paxton Lake benthics and limnetics respectively at a 5% false discovery rate (FDR), suggesting more SNPs were under selection in Paxton Lake benthics than limnetics. Most of the candidate SNPs (68.9%) in Paxton Lake benthics have negative nS_L scores, while the majority of candidate SNPs (81.9%) in Paxton Lake limnetics have positive nS_L scores. This suggests that derived and ancestral alleles are selectively favored by Paxton Lake benthics and limnetics respectively. In addition, there are more SNPs with large negative nS_L scores (nS_L < -2) in Paxton Lake benthics than limnetics (**Fig 2.20**). It indicates selection, especially selection on derived alleles, is more prevalent in Paxton Lake benthics than limnetics.

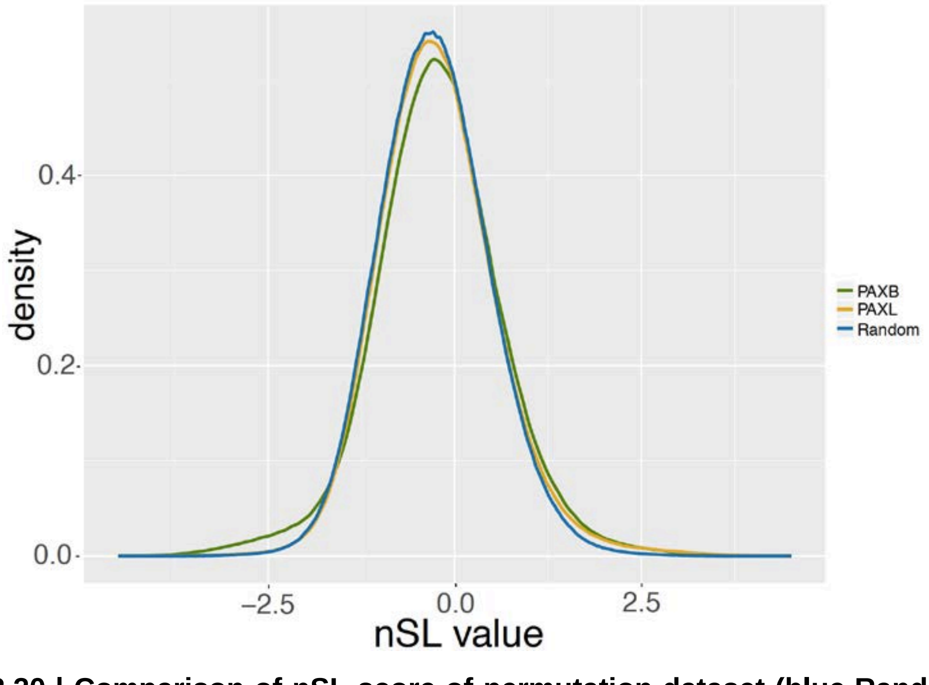

Figure 2.20 | Comparison of nSL score of permutation dataset (blue Random), Paxton Lake benthics (green PAXB), and limnetics (yellow PAXL). The excess of large negative nS_L score in PAXB but not PAXL indicates selection of derived alleles is more prevalent in PAXB than in PAXL.

To determine the origin (derived or ancestral) of selected alleles in divergent regions, I identified SNPs with significant nS_L scores (FDR < 5%) that were located in genomic regions that are highly diverged between Paxton Lake benthics and limnetics (CSS scores, 0.5% of empirical distribution). In Paxton Lake benthics, the majority (88.8%) of selected SNPs that are located in highly divergent regions have negative nS_L scores, indicating derived alleles were selected at these loci. On the other hand, 83.8% of selected SNPs in Paxton Lake limnetics that are located in highly divergent regions have positive nS_L scores, suggesting ancestral alleles were selected at these positions. This suggests selection of derived and ancestral alleles in Paxton Lake benthics and limnetics separately contribute to the genomic divergence of these two species.

2.6 Discussion

Genetic drift is more efficient in fixing or removing variation from the genome when the effective population size (N_e) of the population is small (Hedrick 2005). Therefore, when populations experience a recent reduction in population size (population bottleneck), genetic drift can more easily decrease the genome-wide heterozygosity and the total number of singletons in the genome. Benthics from all four lakes have lower heterozygosity than their limnetic counterparts. Besides, benthics from three of the lakes (Paxton, Priest, Little Quarry Lake) have fewer singletons in the genome than limnetics from the same lake. This suggests that benthics experienced more severe population bottlenecks than limnetics during evolution. Enos Lake benthics have more singletons in the genome than limnetics, which may result from the increased gene flow between species pairs from this lake.

As the linkage disequilibrium (LD) between neutral variants depends on N_e and recombination rate, the recent N_e of populations can be estimated from the extent of LD between neutral variants (Tenesa et al. 2007). Populations with short LD blocks should have large recent N_e. LD between variants on a putatively “neutral” chromosome (chromosome XV) decays more rapidly in marine sticklebacks than in other ecotypes, and LD decays more rapidly in limnetics than in freshwater sticklebacks and benthics. Additionally, benthics have longer LD blocks than all the other three ecotypes, indicating benthics have smaller recent N_e than all other ecotypes. The small recent N_e of benthics may have resulted from: 1) benthics experiencing a more severe population bottleneck than other ecotypes in their evolutionary history; or 2) benthics experiencing a more recent population bottleneck than the other ecotypes. According to the double-invasion hypothesis, benthics invaded the lakes ~1,500 years before limnetics. It is possible that both benthics and limnetics were subject to selection recently due to the competition of resources (character displacement) caused by the invasion of limnetics (Schluter & McPhail 1992). The selection reduced the size of both populations, but the population bottleneck is more severe in benthics than limnetics because the selection is stronger in benthics than in limnetics. The large N_e of marine sticklebacks is consistent with the model that ancestral marine sticklebacks have stable and large N_e, from which freshwater sticklebacks radiates to diverse freshwater systems (Bell & Foster 1994a).

An increased number of hybrids between benthics and limnetics have been found in Enos Lake, suggesting these two species have “collapsed” into one single hybrid swarm (reverse speciation) (Kraak et al 2001, Taylor et al 2006). PCA of benthics and limnetics from all four lakes revealed a shift of the Enos Lake limnetics towards the Enos Lake benthics on the first principal component, which separates benthics from limnetics. In addition, Enos Lake limnetics have a smaller number of singletons in their genome than Enos Lake benthics, which may due to the increased gene flow from benthics to limnetics. Therefore, my analysis provides genetic evidence that the reverse speciation is due to increased gene flow from benthics to limnetics. Interestingly, the samples of Enos Lake limnetics are derived from a group of Enos Lake limnetics transplanted from Enos Lake to a small isolated pond between 1988 and 1989. Previous analysis showed Enos Lake benthics and limnetics collected before 1997 had clear morphological divergence. The authors suggested the reverse speciation might start before 1997 (Taylor et al 2006). My result suggests that the increased gene flow between Enos Lake benthics and limnetics started earlier than 1997, perhaps even before 1988.

As different stages of speciation with or without gene flow have unique patterns of genetic divergence, it is possible to infer the stage of speciation according to the level and distribution of genetic divergence (F_{ST}) in the genome (Seehausen et al 2014). The genome-wide mean F_{ST} between benthics and limnetics from different lakes is similar to the F_{ST} between divergent populations at the late stage of speciation with gene flow and much higher than the F_{ST} between incipient species (Malinsky et al 2015). In addition, the distribution of F_{ST} of benthics and limnetics is similar to the distribution of populations at late stage of speciation with gene flow (Martin et al 2013, Seehausen et al 2014). This suggests benthics and limnetics are at the late stage of speciation with gene flow. The divergence time of divergent populations at the late stage of speciation with gene flow that have similar genome-wide F_{ST} is generally larger than 100,000 years (pied and collared flycatcher: > 300,000 years, Darwin's finches > 900,000 years) (Lamichhaney et al 2015, Nadachowska-Brzyska et al 2013). The ancestors of benthics and limnetics invaded the lakes in the recent 13,000 years (< 13,000 generations) (McPhail 1993, Schluter & McPhail 1992), indicating the ancestors may diverge before invading the lakes. In addition, my result showed sticklebacks, especially benthics and limnetics, have been subject to stronger divergent natural selection than in human. This suggests the large genetic divergence between benthics and limnetics may derive from strong divergent selection and/or pre-existing divergence in ancestral populations.

The genomic regions that are diverged in parallel between populations that were adapted to similar environments repeatedly should be subject to natural selection and contribute to their adaptation (Elmer & Meyer 2011). Benthics and limnetics from different lakes show parallel morphological divergence. My result demonstrated the species pairs from different lakes have high correlation of genetic divergence, indicating the parallel morphological divergence has a genetic basis. In addition, the results revealed that about 15% of genome is diverged among species pairs from all four lakes, suggesting these genomic regions have been subject to divergent natural selection. My result showed derived and ancestral alleles are selectively favored by benthics and limnetics respectively. In addition, more genomic regions have been subject to selection in benthics than limnetics. Therefore, divergence between benthics and limnetics is result from selection of derived and ancestral alleles in these two species, especially selection of derived alleles in benthics. The parallel divergent regions are not evenly distributed throughout the genome. Some chromosomes have substantially more parallel divergent regions than others. Moreover, several parallel divergent regions cluster and form large "islands of genetic divergence".

These “islands of genetic divergence” can facilitate the adaptation of benthics and limnetics as several loci within these “islands” that contribute to local adaptation can be inherited together.

2.7 Materials and Methods

2.7.1 Stickleback samples

2.7.1.1 Benthics and limnetics

Individual fish representing benthics and limnetics from three of the lakes (Paxton, Priest, Little Quarry Lake) were sampled in 2008-2011 and selected based on morphological analyses (discriminant function analysis) to identify individuals most typical/representative of each ecotype. To preserve the Enos Lake limnetics from reverse speciation, 445 individuals of Enos Lake limnetics were introduced to the Murdo Frazer Duck Pond in Murdo-Frazer Park in North Vancouver on September 30, 1988. All these individuals were from 65 families of lab-raised offspring of wild fish. In addition, 150 adult wild Enos Lake limnetics were introduced on May 6, 1989 to supplement earlier introduction. Limnetics were collected from the pond in 1997 and preserved in the lab, and six of them were used as Enos Lake limnetics in this research project. Enos Lake benthics were sampled from Enos Lake in 2008. In addition, seventeen additional individuals of Paxton Lake benthics and limnetics were sampled from Paxton Lake in 2010. In total, 17 individuals each of benthics and limnetics were used in this research project.

2.7.1.2 Marine and freshwater sticklebacks

Marine and freshwater stickleback individuals were collected at 1.5km and 28km from river mouth of Little Campbell River, Canada in 2015. Crosses between male and female individuals were done in the field and embryos were raised in the stickleback fish facility at the Max Planck Institute for Developmental Biology. Sticklebacks were reared in laboratory conditions on Max Planck Campus under 10% seawater (3.5ppt) with daily feeding of both marine and freshwater invertebrates and twice daily 10% water change under Baden-Württemberg Regional Authority permission AZ:35./9185.82-5. Six lab-raised adult individuals each of marine and freshwater sticklebacks were used in this project. Six marine and six freshwater stickleback individuals were collected at 1km and 8km from river mouth of River Tyne, Scotland in 2001 and 2003 separately. In addition, 186 wild caught individuals representing marine and freshwater sticklebacks were sampled by many collaborators across the Northern Hemisphere and collated in Kingsley

Lab, Stanford University, processed, genotyped for sex and females selected (**Appendix Table 3 and 4**).

2.7.2 Whole genome re-sequencing

Genomic DNA was extracted from fin samples following the protocol described previously (Peichel et al 2001). Whole-genome resequencing was performed with different approaches for different sets of stickleback individuals:

1. Six benthics and six limnetics from Paxton lake, Quarry Lake, and Enos Lake, 6 Priest Lake benthics, 3 Priest Lake limnetics, and 6 marine and 6 freshwater sticklebacks from River Tyne were sequenced using Illumina GAIIx with 2×76-bp chemistry at ~13X coverage (**Appendix Table 1 and 3**). Dr. Felicity Jones constructed sequencing libraries and performed whole genome re-sequencing. Three Priest Lake limnetics (PRIL102, PRIL108, PRIL112) were sequenced with 2×150bp chemistry on an Illumina HiSeq 3000. Sequencing libraries were constructed following Illumina TruSeq sequencing library construction protocol with homemade reagents on TECAN liquid handling machine. All three individuals were barcoded with Illumina TruSeq adapters and sequenced with samples of other projects in one lane to reach ~20X coverage (**Appendix Table 1**). My colleague, Ms. Vrinda Venu, constructed the sequencing libraries. Sequencing team of Max Planck Institute for Developmental Biology performed sequencing.
2. Seventeen individuals each of Paxton Lake benthics and limnetics were sequenced using Illumina HiSeq 3000 with 2×150bp chemistry. Sequencing libraries were constructed following Illumina TruSeq sequencing library construction protocol with homemade reagents on TECAN liquid handling machine. Seventeen individuals were barcoded with Illumina TruSeq adapters and sequenced in on lane to reach ~20X coverage (**Appendix Table 2**). Ms. Vrinda Venu, constructed the sequencing libraries. Sequencing team of Max Planck Institute for Developmental Biology performed sequencing.
3. Six individuals each of marine and freshwater ecotypes from Little Campbell River were sequenced using Illumina HiSeq 3000 with 2×150bp chemistry. Sequencing libraries were constructed following Illumina TruSeq sequencing library construction protocol with homemade reagents. Twelve individuals were barcoded with Illumina TruSeq adapter and sequenced with samples of other projects in on lane of HiSeq 3000 with to reach ~15X coverage (**Appendix Table 3**). Dr. Jukka-Pekka Verta constructed the sequencing library. Sequencing team of Max Planck Institute of Developmental Biology performed sequencing.

4. One hundred and eighty nine individuals of marine and freshwater sticklebacks were sequenced at ~5X coverage (**Appendix Table 4**). DNA for genome sequencing was shipped to Broad Institute for whole genome sequencing with 2x100bp chemistry on an Illumina HiSeq 2000

2.7.3 SNP calling and filtering

2.7.3.1 SNP calling

The sequencing reads of stickleback individuals were aligned to stickleback reference sequence (Broad S1) (Jones et al 2012b) using Burrows-Whleeler Aligner (BWA) v0.7.10-r789 (Li & Durbin 2010) with BWA `mem` function. Custom pipeline of SNP detection following GATK best practices was performed:

- Sort and index SAM file using SortSam program of Picard Tools v1.128 (https://broadinstitute.github.io/picard/).
- Remove PCR duplicates using MarkDuplicate program of Picard Tools.
- Local realignment of reads around Indels using IndelRealigner program of Genome Analysis Toolkit (GATK) (McKenna et al 2010) v3.4.
- Base quality recalibration of sequencing reads using BaseRecalibrator program of GATK v3.4. The reference dataset of known SNPs was generated from previously published SNP dataset of 21 marine and freshwater sticklebacks (Jones et al 2012b). Only SNP sites have 8 reads support at all individuals were retained.
- Coverage of each individual was evaluated using DepthOfCoverage program of GATK v3.4.
- SNP variants were identified using Haplotypecaller program of GATK v3.4

2.7.3.2 SNP filtering and validation

SNP filtering was performed using GATK Variant Quality Recalibration (VQSR) pipeline. Firstly, variant quality scores were recalibrated using a reference dataset of known SNPs. Due to the lack of “golden” quality reference variant set of sticklebacks, the reference dataset was generated by filtering SNP dataset of 206 marine and freshwater stickleback raw SNP calling dataset using Hard Filtering pipeline of GATK with parameters: `QD < 2.00 || FS > 60.000 || MQ < 50.00 || MQRankSum < -12.500 || ReadPosRankSum < -8.000`. Secondly, four sensitivity tranches (95%, 99%, 99.5%, 99.9%) of variant quality were calculated according to the known

SNPs in reference dataset. Lastly, SNPs were filtered with 99.9% sensitivity tranche and only bi-allelic SNPs were kept in the dataset.

To estimate the error rate of the SNP calling, I validated the SNPs in the cross-lake benthics and limnetics SNP dataset by Sanger sequencing. PCR primers were designed for 94 randomly selected SNPs in the cross-lake benthics and limnetics variant dataset (**Appendix Table 10**). Genomic regions containing these SNPs were amplified and Sanger sequenced for all 64 individuals of cross-lake benthics and limnetics. Of 94 genomic regions being analyzed, 74 of them were successfully amplified and sequenced. Most SNPs being tested represent true SNPs (69/74, 93.2%).

2.7.3.3 Phasing

SNPs in the Paxton Lake benthics and limnetics variants dataset was phased using the read aware phasing algorithm (Delaneau et al 2013) implemented in SHAPEIT v2.r837 (Delaneau et al 2013). Read aware phasing method identifies phase informative reads (PIR) which span at least two heterozygous sites and uses these reads to improve the accuracy of phasing. Firstly, phase informative reads were extracted from the alignment files of 3 individuals each of Paxton Lake benthics and limnetics with similar and high coverage (~20X) in the dataset (PAXB105, PAXB115, PAXB119, PAXL128, PAXL139, PAXL150) using extractPIRs tool of SHAPEIT. Secondly, SNPs were phased using phase informative reads and previously published stickleback genetic map (Roesti et al 2013) as guidance. SNPs in the cross-lake benthics and limnetics variants dataset were also phased using read aware phasing algorithm. Phasing was performed with phase informative reads of benthics and limnetics having high coverage in the dataset (PAXB05, PAXB07, PAXL01, PAXL14, PRIB07, PRIB15, PAXL102, PRIL16, QRYB01, QRYB13, QRYL05, QRYL10, ENSB08, ENSB12, ENSL24, ENSL25) and previously published genetic map (Roesti et al 2013) as guidance using SHAPEIT.

2.7.4 Genomic composition of benthics and limnetics

2.7.4.1 Principal Component Analysis (PCA)

PCA of benthics and limnetics from all four lakes was performed using smartpca program v13050 using genome-wide SNPs (Patterson et al 2006). A total of 6,134,540 SNPs were used in the analysis after filtering by smartpca program. SNPs with high degree of linkage disequilibrium (LD) were removed using the LD correction function of smartpca program with option “`nsnpldregress 2`”. PCA of Paxton Lake benthics and limnetics using the smartpca program in EIGENSOFT package v13050 using with default setting. In total, 131,132 SNPs were used for the analysis after filtering by smartpca program. SNPs with high degree of LD were removed using the LD correction function of smartpca program with option "`nsnpldregress 2`". The results of PCA were plotted using custom R script.

2.7.4.2 *Genomic diversity*

Average heterozygosity (measured by 2pq) was calculated for each population using custom Python script. The genome-wide average of heterozygosity at each SNP was calculated as the average heterozygosity of each population. Genome-wide nucleotide diversity (π) of each population was calculated using VCFtools v0.1.14 (Danecek et al 2011). Singleton SNPs of each individual were calculated using VCFtools v0.1.14. To eliminate the effect of missing data and depth variation, only sites with no missing SNP call in all studied individuals were used for the calculation of singleton SNPs. The results were plotted using custom R script. Genome-wide genetic divergence between different stickleback populations was estimated by F_{ST} using VCFtools v0.1.14.

2.7.4.3 *Allele frequency spectrum*

To infer the unfolded allele frequency spectrums, the ancestral allele at each SNP site was determined as the most frequent allele of marine individuals from Little Campbell River and River Tyne using custom Python script, and the derived allele was determined as the alternative allele. Derived allele frequency for each SNP site was calculated using VCFtools v0.1.14. The allele frequency spectrum of benthics and limnetics from each lake were plotted using custom R script. Two-dimensional site frequency spectrum of Paxton Lake benthics and limnetics was generated using Paxton Lake benthics and limnetics variant dataset. The ancestral allele was determined as the most frequent allele in marine individuals from Little Campbell River and River Tyne. The two-dimensional frequency spectrum was plotting using δaδi package (Gutenkunst et al 2009a) v1.7.0.

2.7.4.4 *Comparison of genome-wide mean F_{ST} and extreme allele frequency difference*

Comparison of genome-wide mean F_{ST} and extreme allele frequency difference was performed using Benthic, limnetic and global stickleback SNP dataset as well as human SNP dataset downloaded from 1000 Genomes project website (http://www.internationalgenome.org/data#download). Fourteen human populations were selected for the analysis to achieve better representation of human genetic divergence (**Table 2.3**). To remove the effect of sample size variations of stickleback and human, 6 individuals were randomly selected for human populations with sample size larger than 6. Ancestral allele of sticklebacks was determined according to the most frequent allele of all marine individuals in the dataset. Ancestral allele of human was assigned by 1000 Genomes consortium. Derived allele frequency at each variation site of stickleback and human individuals was calculated using VCFtools V0.1.14. Pairwise extreme allele frequency difference (95% percentile, 99% percentile, maximum) of stickleback and human populations was calculated using custom Python script. Pairwise genome-wide F_{ST} of stickleback and human populations was calculated using VCFtools V0.1.14. Results were plotted using custom R script.

2.7.4.5 Linkage disequilibrium

Linkage disequilibrium (LD) was estimated for putative “neutral” chromosome XV using PLINK v1.90 (Purcell et al 2007). LD was calculated for benthic and limnetic sticklebacks from all four lakes as well as marine and freshwater ecotype of Little Campbell River within 100kb window with option “`--ld-window 99999 --ld-window-kb 100 –ld-window-r2 0`”. The plot of LD decay used r2 measure of LD, and show averages within 1000bp windows using custom R script.

2.7.5 Genomic pattern of adaptive divergence of benthics and limnetics

2.7.5.1 Cluster separation score (CSS)

CSS scores were calculated by subtracting the mean of π between two individuals from different populations by the mean of π between two individuals from the same populations in sliding windows (size: 2,500bp; step: 500bp) using the previously described equation (Jones et al 2012b) with custom Python script. The nucleotide diversity (π) was calculated for all possible pairs of two benthic or limnetic individuals in the dataset for each window using VCFtools V0.1.14. Genome-wide distributions of CSS scores of benthics and limnetics from all four lake as well as Paxton Lake benthics and limnetics were plotted using custom R script.

Large permutation test was performed to determine how many regions were significantly deviated from neutral expectation. I want to calculate all possible combinations of 24 individuals each of cross-lake benthics and limnetics or 23 individuals each of Paxton Lake benthics and limnetics for each window. However, all possible combinations for both dataset are extremely large [1.61×10^{13} combinations (~5 million CPU hours) for each window of cross-lake benthics and limnetics; 1.214×10^{7} combinations (~4 CPU hours) for each window of Paxton Lake benthics and limnetics], which is impossible to calculate for all 926,407 windows in the genome. Thus, I determined to calculate CSS scores for 1 million combinations for each window. For cross-lake benthics and limnetics, I calculated CSS scores for 1 million random combinations in dividing into two groups of 24 and 24 individuals at all 926,407 windows using custom Python script. *P*-values were calculated using custom C++ script with the resulting 1 million CSS scores at each window. For Paxton Lake benthics and limnetics, CSS scores were calculated for 1 million random combinations in dividing into two groups of 23 and 23 individuals at all 926,407 windows using custom Python script. *P*-values were calculated using custom C++ script with the resulting 1 million CSS scores at each window.

2.7.5.2 *SweepFinder2*

SweepFinder2 (DeGiorgio et al 2016) v1.0 was used to detect complete selective sweep in the genomes of benthics and limnetics. The ancestral allele at each SNP was determined according to the most frequent allele of marine ecotypes from Little Campbell River and River Tyne and neutral SFS was calculated using all SNPs in the genome. Genetic distance between SNPs was calculated using previously published genetic map of stickleback (Roesti et al 2013). For the sweep detection of cross-lake benthics and limnetics, benthics or limnetics from all four lakes were combined for the analysis. In total, 6,637,116 and 7,601,856 SNPs of benthics and limnetics were input into SweepFinder2 separately after filtering according to the software's requirement. Selective sweeps were detected in non-overlapping windows (2,500 bp) with default settings of SweepFinder 2. The result was plotted using custom R script.

After filtering, 9,864,613 and 8,374,445 SNPs were input into SweepFinder2 separately to detect selective sweeps in the genomes of Paxton Lake benthics and limnetics in non-overlapping windows (2,500 bp). The genomic distributions of CLR were plotted using custom R script.

2.7.5.3 *nS_L*

Genomic regions under selection in Paxton Lake benthics and limnetics were identified using nS_L (Ferrer-Admetlla et al 2014) with default setting. SHAPEIT phased SNPs of Paxton Lake benthics and limnetics were polarized using the marine ecotypes from Little Campbell River and River Tyne as outgroup. The input files for nS_L program were generated using custom Python scripts. The nS_L runs were performed for Paxton Lake benthics and limnetics separately and each chromosome independently.

I performed permutation tests to evaluate the significance of nS_L scores. Firstly, I randomly selected 1,000 regions with length of 1Mb from the genome. All the selected regions have to be at least 1Mb away from the end of chromosome. Secondly, for each of 1,000 regions, I randomly selected 46 haplotypes (23 individuals) 100 times and obtained a dataset. Lastly, I calculated nS_L score for all 100,000 datasets. In the end, 536,396,550 nS_L scores were obtained and combined as the null distribution. I used this null distribution to calculate *P*-value for each empirical nS_L score of Paxton Lake benthics and limnetics using FastPval (Li et al 2010). Results were plotted using custom R scripts. SNPs with false discovery rate (FDR) less than 5% were identified as significantly deviated from neutral expectation.

2.7.5.4 Derived and ancestral haplotypes at divergent regions

To determine whether divergent regions of benthics and limnetics carry derived or ancestral haplotypes, F_{ST} between benthics or limnetics and marine sticklebacks from Little Campbell River and River Tyne at each divergent region was calculated using VCFtools V0.1.14. The distribution of F_{ST} scores was plotted using custom R script.

3 FUNCTIONS AND SOURCES OF ADAPTIVE GENETIC VARIATION IN BENTHICS AND LIMNETICS

3.1 Background and Aims

Identifying and analyzing adaptive loci in various organisms provides insight into how natural selection shapes the genome and individual traits during evolution (Wolf & Ellegren 2017). Furthermore, studying the origin of adaptive variation helps elucidate how genetic polymorphisms are maintained within a natural population (Barrett & Schluter 2008b).

Researchers have started to study the adaptive loci of benthic and limnetic sticklebacks. Using benthic and limnetic crosses, researchers have identified several QTLs underlying morphological trait differences (Arnegard et al 2014, Conte et al 2015). These large-scale QTL mapping analyses provided valuable insights into benthic and limnetic adaptation: 1) several loci with small to moderate phenotypic effect are required in benthic and limnetic divergence (Arnegard et al 2014); 2) adaptation of benthics and limnetics is a complicated process involving several interacting phenotypic traits regulated by multiple genomic loci (Arnegard et al 2014); and 3) nearly half of the genomic regions have been repeatedly used by benthics and limnetics during their adaptation in different lakes (Conte et al 2015). However, due to relatively small clutch size, QTL mapping in sticklebacks is typically low resolution; detection is limited to loci with large effects (Berner & Salzburger 2015). Thus, these two QTL mapping studies have not identified the genes or regulatory factors contributing to benthic-limnetic adaptation. In addition, QTL mapping studies usually focus on traits that are easy to manipulate and measure (Savolainen et al 2013). The genomic loci regulating adaptive traits that have subtle or invisible phenotypic divergence (i.e. blood circulation) between populations cannot be resolved by QTL mapping. A higher-resolution approach is needed to identify the loci which control diverse adaptive traits.

Theoretical studies have shown that rapid adaptation likely arose from selection of standing genetic variation, as the new beneficial mutations are not immediately available for selection in a population when the environment changes (Barrett & Schluter 2008b). Genetic and genomic studies show that during their adaptation to the new environment, freshwater sticklebacks tend to use genetic variations which were present at low frequency in marine sticklebacks (the “transporter” hypothesis) (Colosimo et al 2005, Jones et al 2012b, Schluter & Conte 2009). Selection on *de novo* mutations also contributes to freshwater stickleback adaptation as seen at the *Pitx1* enhancer, where repeated *de novo* deletions resulted in pelvic spine reduction (Chan et al 2010). Sympatric benthic and limnetic stickleback pairs are rare, and show substantial phenotypic divergence due to adaptation to their own environmental niches (McPhail 1994). Thus, these two species have made use of both standing genetic variation and *de novo* mutation during adaptation.

In this chapter,

- I identify and characterize the adaptive loci of Paxton Lake benthics and limnetics as well as those of benthics and limnetics from all four lakes.
- I determine the source of adaptive genetic variation in the genome of benthics and limnetics.

3.2 Adaptive loci of Paxton Lake benthics and limnetics

Natural selection increases between-population divergence at beneficial genomic regions in populations living in different environmental niches (see **Section 1.2.3**) (Vitti et al 2013). Therefore, population divergence is commonly used to detect adaptive loci in the genome (Holsinger & Weir 2009b), and the divergent regions of Paxton Lake benthics and limnetics identified in previous chapter (see **Section 2.5.1**) are likely to contribute to their adaptation. However, “divergent hitchhiking” (see **Section 1.4.2**) can result in large genomic regions with elevated divergence (islands of genetic divergence) (Nosil et al 2009a). These regions contain numerous neutral alleles that do not contribute to adaptation. As the genomic regions under divergent natural selection show higher divergence and stronger selection signatures than neutral regions, it is possible to identify adaptive loci by looking for regions with high genetic divergence and other signatures of selection. Therefore, I identified adaptive loci by looking for highly divergent genomic windows (2,500bp; step: 500bp; CSS: top 0.5%) that show strong signals of selection in CLR analysis (top 0.5%) or contain at least one SNP with a significant nS_L score (FDR < 5%) in *both* Paxton Lake benthics and limnetics. In total, 465 windows (131 genomic regions) covering 518,870bp of the genome (0.11%) were identified as adaptive regions of Paxton benthics and limnetics (**Appendix Table 11**). More than half of the genome is diverged between Paxton Lake benthics and limnetics (see **Section 2.5.1**). However, only a small proportion of these divergent regions have been subject to divergent selection in these two species. This can be attributed to "divergent hitchhiking" (see **Section 1.4.2**), by which numerous neutral alleles can be carried to high frequency by sweeps of nearby beneficial alleles (Nosil et al 2009a). Nonetheless, neutral alleles do not show signatures of selection. Therefore, combining several statistics greatly improves the detection of adaptive loci in Paxton Lake benthics and limnetics.

The genomic regions carrying divergent derived haplotypes that were selected in both Paxton Lake benthics and limnetics are important for each population's unique adaptation. Therefore, I identified these regions by looking for adaptive regions of Paxton Lake benthics and limnetics that have high divergence ($F_{ST} > 0.5$) between marine sticklebacks (marine ecotypes from Little Campbell River and River Tyne) and both Paxton Lake benthics and limnetics. There are 11 adaptive regions on chromosomes IV, VII, and VIII where divergent derived haplotypes were selected in Paxton Lake benthics and limnetics (**Table 3.1**). These regions overlap with 5 genes, two of which (*SCUBE1*, *COL24A1*) have important functions in vertebrate (especially zebrafish) development. *Signal peptide-CUB domain-EGF-related-1* (*SCUBE1*) regulates bone morphogenetic protein (BMP) signaling during primitive hematopoiesis in zebrafish (*Danio rerio*) (Tsao et al 2013). Knockdown of *SCUBE1* caused the anterior-posterior axis to be shortened in zebrafish (Johnson et al 2012). Anterior-posterior axis length is a morphological trait that differs between benthic and limnetic sticklebacks (Schluter & McPhail 1992). *Collagen type XXIV alpha1* (*COL24A1*) is associated with osteoblast differentiation and bone formation in mouse

(Matsuo et al 2008) and in regeneration of fin skeleton in zebrafish (Duran et al 2015).

One of the adaptive regions where divergent derived haplotypes were selected in Paxton Lake benthics and limnetics overlaps with an intergenic region flanked by two genes (*AR* and *MSNA*) known to regulate important phenotypic traits in zebrafish. *Androgen receptor* (*AR*) encodes the cytosolic receptors of androgen ligands that influence male courtship behavior in zebrafish (Yong et al 2017). Upon *AR* knockdown, male zebrafish mated with females significantly less often. *Moesin a* (*MSNA*) plays an important role in maintaining apical and basal cell polarity within intersegmental vessels in the zebrafish embryo (Wang et al 2010).

Another adaptive region carrying divergent derived haplotypes overlaps with a protein coding gene (ENSGACG00000007263) which is the ortholog of zebrafish *Phosphodiesterase 4B, cAMP-specific a* (*PDE4BA*) gene. Interestingly, a genomic region upstream of this gene was highly divergent between global marine and freshwater sticklebacks (Jones et al 2012b). This suggests that alternate haplotypes carried by marine and freshwater sticklebacks confer selective advantages in their respective marine and freshwater environments. Benthics and limnetics carry divergent derived haplotypes at this region. As both benthics and limnetics live in freshwater environments, these two alternative haplotypes should confer a fitness advantage in their respective environmental niches.

Table 3.1 Adaptive regions where divergent derived haplotypes were selected in Paxton Lake benthics and limnetics

No.	Chromosome	Start	End	Ensembl gene ID	Gene name	Ensembl gene ID (flanking gene)
1	chrIV	24,120,001	24,127,000	ENSGACG00000019325	*SCUBE1*	
2	chrIV	24,155,001	24,157,500			
3	chrVII	17,149,501	17,153,000	intergenic region		*AR* *MSNA*
4	chrVIII	7,053,501	7,057,500	ENSGACG00000006637	*COL24A1*	
5	chrVIII	7,082,501	7,086,000			
6	chrVIII	7,090,501	7,094,500			
7	chrVIII	7,101,501	7,106,500			
8	chrVIII	7,109,501	7,113,500			
9	chrVIII	7,946,501	7,951,000	ENSGACG00000007122	*RAVER2*	
10	chrVIII	8,208,501	8,212,000	ENSGACG00000007263	*PDE4BA-like*	
11	chrVIII	8,330,501	8,335,000	ENSGACG00000007270		

Although derived alleles were selected in Paxton Lake benthics at most of the adaptive regions, there are three adaptive regions where derived alleles were selected only in Paxton Lake limnetics (chrVIII: 8,369,501-8,374,500; chrVIII: 8,381,501-8,386,000; chrUn: 1,481,501-1,486,000). The two adaptive regions on chromosome VIII overlap with *Hemicentin 1* (*HMCN1*). *HMCN1* is a large gene spanning over 62kb (chrVIII: 8,358,318-8,421,177), and the two adaptive regions overlap with a small section of the gene (15.1%). To comprehensively study the selective signature of *HMCN1*, I investigated its genotype in Paxton Lake benthics and limnetics. Interestingly, Paxton Lake benthics and limnetics carry different derived haplotypes at *HMCN1* (F_{ST} > 0.5)(**Fig. 3.1**). Both Paxton Lake benthics and limnetics contain several missense mutations at *HMCN1*, suggesting its function may diverge in these two species. *HMCN1* regulates medial fin development in zebrafish (Carney et al 2010, Westcot et al 2015). Zebrafish knockdown mutants of *HMCN1* generate embryos of fin blister (Westcot et al 2015). This suggests *HMCN1* may be critical for the adaptation of Paxton Lake benthics and limnetics.

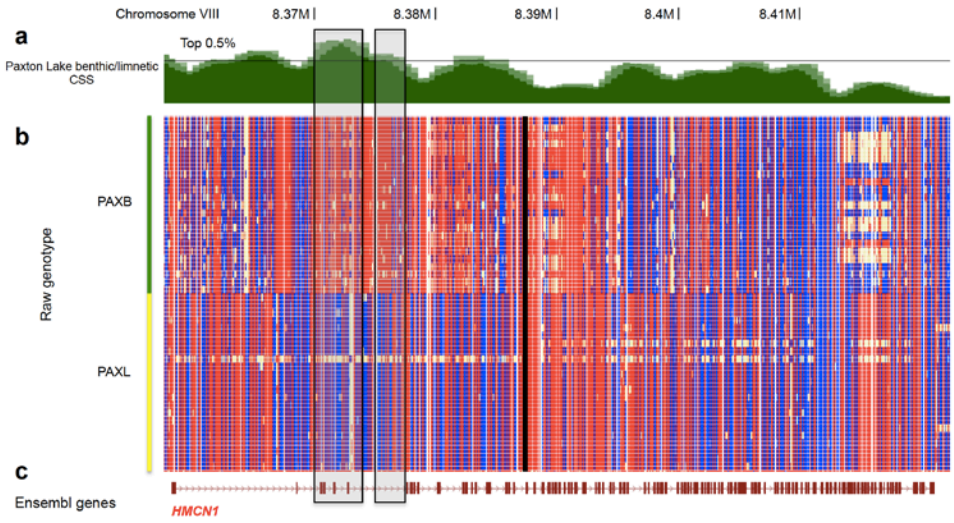

Figure 3.1 | Visual genotype for Paxton Lake benthics (PAXB) and limnetics (PAXL) at *HMCN1*. a, CSS scores of Paxton Lake benthics and limnetics. **b,** Visual genotype for Paxton Lake benthics and limnetics. Red represents the most frequent allele in the marine ecotype from Little Campbell River and River Tyne (ancestral alleles), blue represents alternative (derived) alleles, and yellow, heterozygous alleles. **c,** Ensembl gene model. The two adaptive regions where Paxton Lake limnetics are carrying the derived allele are shown as vertical shaded boxes.

The genomic regions where different derived haplotypes were selected in Paxton Lake Benthics and limnetics played a critical role in their adaptation to their unique habitats. Genetic divergence at these loci has been maintained despite the homogenizing effects of ongoing gene flow, suggesting that the alternative haplotypes confer an adaptive advantage to the respective ecotypes. Genes residing in or adjacent to adaptive regions where divergent derived haplotypes were selected in Paxton Lake benthics and limnetics have been shown to regulate bone (*SCUBE1*, *COL24A1*), fin (*HMCN1*), and blood vessel development (*MSNA*) as well as male courtship behavior (*AR*) in zebrafish. Selection of these genes during adaptive divergence of Paxton Lake benthics and limnetics might contribute to their divergent body size and body shape, and furthermore to the reproductive isolation of these two species.

3.3 Adaptive loci of benthics and limnetics

3.3.1 Adaptive loci of benthics and limnetics where both benthics and limnetics have been subject to selection (“Strongly adaptive loci”)

Benthics and limnetics from different lakes show parallel morphological divergence, which is strong evidence of natural selection. The study of genomic patterns of genetic divergence demonstrated that there were genomic regions consistently diverging among benthics and limnetics from all four lakes (see **Section 2.4.2**). These regions contributed to the adaptation of benthics and limnetics and should be subject to positive selection, as it is unlikely that genetic drift would fix the same alleles in benthics or limnetics from all four lakes. To identify adaptive loci in benthics and limnetics, I examined the 465 adaptive windows of Paxton Lake benthics and limnetics previously identified (**see Section 3.2.1**) as highly diverged in benthics and limnetics from all four lakes (CSS: top 0.5%). In total, 237 out of 465 adaptive windows (50.9%) of Paxton Lake benthics and limnetics have extreme CSS scores in benthics and limnetics from all four lakes, indicating these regions contributed to the parallel adaptation of benthics and limnetics. This is similar to the previous estimation (48.8%) of QTL reuse in the adaptation of benthics and limnetics from Paxton and Priest Lake (Conte et al 2015). After concatenating overlapping windows, 77 adaptive genomic regions covering 284,923bp of the genome (0.06%) were recovered, which I refer to as "strongly adaptive regions". (**Appendix Table 12**). In total, 33 genes lie within or overlap with these "strongly adaptive regions" (**Table 3.2**). These genes regulate several important biological processes in zebrafish, including body development (*USP25*, *MED13A*, *WNT5A*), live morphogenesis (*NAV3*), eye development (*EBNA1BP2*, *OPA1, DNMT3BB.2*), lipid metabolism (*GDPD5A*, *B4GALNT1A*) and cardiovascular development (*OPA1*, *TNNT2A*) (**Table 3.2**). In addition, some of the "strongly adaptive regions" are located in entirely intergenic regions. The 35 genes proximal to these intergenic regions regulate several biological processes in zebrafish, including body (*LARP7*), eye (*AUTS2A*), and cardiovascular development (*LARP7*, *MSNA*) as well as lipid metabolism (*ACOX3*) (**Table 3.3**). These processes may be important for the speciation and adaptation of benthics and limnetics.

There is parallel divergence in body size of benthics and limnetics from different lakes. Benthics have a larger body than limnetics (McPhail 1994). Both benthic and limnetic females prefer to mate with conspecific males having body sizes similar to their own (Boughman et al 2005). This suggests body size may contribute to mating preference and thus reproductive isolation of these two species. *Wingless-type MMTV integration site family, member 5a* (*WNT5A*) is one of the key regulators of osteoblast formation (Maeda et al 2012). *WNT5A* can both positively and negatively regulate Wnt/β-catenin signaling in the mouse embryo (van Amerongen et al 2012). Mouse knockout and zebrafish knockdown mutants of *WNT5A* have both a smaller body size and a shorter anterior-posterior axis than wild-type individuals (Huang et al 2014, van Amerongen et al 2012). *Ubiquitin specific peptidase 25* (*USP25*) is a positive regulator of Wnt/β-catenin signaling (Xu et al 2017). Knocking down *USP25* gene in zebrafish decreased the length of the anterior-posterior axis (Tse 2017). Therefore, parallel divergence at these two loci in benthics and limnetics may be due to strong selection of their body size during adaptation, and may further contribute to mating preference.

Intriguingly, three genes (*OPA1, DNMT3BB.2, EBNA1BP2*) regulating eye development have been subject to divergent selection in benthics and limnetics from different lakes. *Optic atrophy 1* (*OPA1*) plays an important role in human eye vision. Mutations in *OPA1* lead to optic atrophy and eventually blindness (Alexander et al 2000, Delettre et al 2000). In addition, zebrafish *OPA1* knockdown mutants have decreased eye size compared to wild-type individuals (Rahn et al 2013). *DNA (cytosine-5-)-methyltransferase 3 beta, duplicate b.2* (*DNMT3BB.2*) is one of the major genes regulating eye development in zebrafish. In zebrafish *DNMT3BB.2* knockdown mutants, retinal development is disrupted, and the retina pigmented epithelium ventral region is absent (Rai et al 2010). *EBNA1 binding protein 2* (*EBNA1BP2*) is also involved in zebrafish eye development. Zebrafish knockdown mutants of *EBNA1BP2* have smaller eye size than wild-type individuals (Amsterdam et al 2004). This suggests that some key genes regulating fish eye development have been divergently selected in benthics and limnetics. Divergence of color vision in benthics and limnetics is critical for prezygotic reproductive isolation since limnetic females distinguish benthic from limnetic males by their nuptial coloration (Boughman et al 2005). Because the genes that regulate eye development in benthics and limnetics contribute to their mating preference, they are good candidates for selection.

Table 3.2 Functions of genes overlapping with "strongly adaptive regions" of benthics and limnetics from all four lakes

No.	Ensembl Gene ID	Gene Name	Zebrafish Gene Ontology Annotation*	Zebrafish Knockdown Phenotype	Reference
1	ENSGACG00000019472	*NAV3*	liver morphogenesis, pancreas development	reduced liver size, impaired development of pancreas and swim bladder	(Klein et al 2011)
2	ENSGACG00000000641	si:dkey-28n18.9			
3	ENSGACG00000000833	*COL7A1*	*epidermis development, extracellular matrix organization (human)*	*mutation in COL7A1 in human causes Epidermolysis bullosa dystrophica*	(Dang & Murrell 2008)
4	ENSGACG00000020027	*RASGRP2*	intracellular signal transduction		
5	ENSGACG00000020030	*NRXN2 (1 of many)*	*signal transduction, neurotransmitter secretion (human)*		
6	ENSGACG00000020078	*MPDU1B*	dolichol-linked oligosaccharide biosynthetic process		
7	ENSGACG00000020152	*SERPINH1A*	collagen fibril organization		
8	ENSGACG00000020153	*GDPD5A*	lipid metabolic process		
9	ENSGACG00000020155	*USP25*	cranial skeletal system development, dorsal/ventral pattern formation	malformation of the facial skeleton, dorsalization, small and short body, unshaped eye	(Tse 2017, Tse et al 2009)
10	ENSGACG00000020213	*MED13A*	regulation of transciption from RNA polymerase II promoter	multiple tail bud phenotype	(Lin et al 2007)
11	ENSGACG00000020236	*TSPOAP1-like*	*neurotranmitter secretion, glutamate secretion (human)*		
12	ENSGACG00000020239	*MTMR4*	dephosphorylation, transforming growth factor beta receptor signaling pathway		
13	ENSGACG00000020240	*CA4A*	one-carbon metabolic process	decrease in Na$^+$ accumulation in H+-ATPase/mitochondrion-rich cells (H-MRCs)	(Ito et al 2013)
14	ENSGACG00000020335	*AMER1*-like	anatomical structure development,	decreased eye size, malformed head	(Major et al 2007)

			regulation of canonical *Wnt* signaling pathway		
15	ENSGACG00000006695	*ZNHIT6*	*ribosome biogenesis, protein oligomerization (human)*		
16	ENSGACG00000006802	*EBNA1BP2*	ribosomal large subunit biogenesis, rRNA processing	small head and eyes, thin body and less pigment	(Amsterdam et al 2004)
17	ENSGACG00000007249	*SGIP1A*			
18	ENSGACG00000007263	*PDE4BA*	signal transduction		
19	ENSGACG00000007270				
20	ENSGACG00000009278	*OPA1*	chordate embryonic development, mitochondrial fission, ventricular cardiac muscle cell development	disrupted blood circulation, decreased eye size, decreased heart size	(Rahn et al 2013)
21	ENSGACG00000009295	*ATP13A3*	cation transport		
22	ENSGACG00000011015	*SOCS3A*	regeneration, cytokine-mediated signaling pathway, posterior lateral line neuromast hair cell development, retina morphogenesis	decreased number of posterior lateral line neuromasts	(Liang et al 2012)
23	ENSGACG00000008972	*B4GALNT1A*	lipid glycosylation		
24	ENSGACG00000009747	*TNNT2A*	artery development, blood circulation, heart contraction	decreased blood circulation rate, heart contraction absent	
25	ENSGACG00000010123	*CACNA2D3 (1 of many)*	calcium ion transport, cardiac conduction (human)		
26	ENSGACG00000010153	*WNT5A*	neuronal differentiation, pronephros development, *Wnt* signaling pathway	decreased eye size, decreased whole organism anatomical axis length, dilated pronephric glomerulus	(Huang et al 2014)
27	ENSGACG00000010256	*MAPRE1B*			

28	ENSGACG00000010262	*DNMT3BB.2*	neurogenesis, eye photoreceptor cell development, retina layer formation	decreased brain and head size, disorganized retina, camera-type eye photoreceptor cell differentiation disrupted	(Rai et al 2010)
29	ENSGACG00000010294	*NOL4L (1 of many)*			
30	ENSGACG00000010479	*OPRL1*	neuropeptide signaling pathway		
31	ENSGACG00000010484				
32	ENSGACG00000010687	*SPRYD3 (1 of many)*			
33	ENSGACG00000008407	*TACC1*	cell proliferation		

Note: Zebrafish and human gene ontology (GO) annotations were obtained from the Amigo database (The Gene Ontology 2017).

Table 3.3 Functions of genes flanking the "strongly adaptive regions" of benthics and limnetics

No.	Ensembl Gene ID	Gene Name	Zebrafish Gene Ontology Annotation*	Zebrafish Knockdown Phenotype	Reference
1	ENSGACG00000009072	*GRIK4*	Ion transport		
2	ENSGACG00000009080				
3	ENSGACG00000019001	*PIM3*	Negative regulation of apoptotic process, protein phosphorylation, regulation of mitotic cell cycle		
4	ENSGACG00000019005	*CRELD2*			
5	ENSGACG00000019011	*FAM19A5A*			
6	ENSGACG00000019014	*TBC1D22A*	activation of GTPase activity, intracellular protein transport, regulation of vesicle fusion		
7	ENSGACG00000018065	*CDK2AP2*	*phosphorylation, regulation of microtubule cytoskeleton organization, regulation of stem cell division (human)*		
8	ENSGACG00000018066	si:dkey-27p18.2			
9	ENSGACG00000018706				
10	ENSGACG00000018707	*RMB20-like*	*mRNA processing, heart development, regulation of RNA splicing (human)*		
11	ENSGACG00000000758	*NAT16-like*	protein acetylation		
12	ENSGACG00000000759	*ACOX3*	fatty acid metabolic process, oxidation-reduction process		
13	ENSGACG00000019605	*ETNPPL*	embryonic hemopoiesis	decreased number of blood cells, embryonic hemopoiesis disrupted	(Eckfeldt et al 2005)
14	ENSGACG00000019618	*LARP7*	regulation of cardiac muscle cell proliferation, cardiac muscle tissue regeneration	animal organ development disrupted, brain degenerate, embryo development disrupted	(Barboric et al 2009, Matrone et al 2015)
15	ENSGACG00000019921				
16	ENSGACG00000019922	*EFNB3A*	axon guidance, ephrin receptor signaling pathway		
17	ENSGACG00000020154	*NRIP1B*	regulation of transcription from RNA polymerase II promoter		
18	ENSGACG00000020158	*SIM2*	regulation of transcription, DNA-templated		

19	ENSGACG00000020159	*HLCS*	cellular protein modification process		
20	ENSGACG00000020210	*AUTS2A*	chordate embryonic development, embryonic viscerocranium morphogenesis, forebrain neuron development	decreased eye, fin, and head size	(Beunders et al 2013, Oksenberg et al 2013)
21	ENSGACG00000020211	*si:ch211-14a17.7*			
22	ENSGACG00000020235	*SSC4D*	*receptor-mediated endocytosis (human)*		
23	ENSGACG00000020237	*SUPT4H1*	Chromatin organization, positive regulation of DNA-templated transcription, elongation		
24	ENSGACG00000020238	*HPDA*	aromatic amino acid family metabolic process, L-phenylalanine catabolic process		
25	ENSGACG00000020241	*GUSB*	carbohydrate metabolic process		
26	ENSGACG00000020259	*CENPV*	metabolic process		
27	ENSGACG00000020260	*NCOR1-like*	Anterior/posterior pattern specification, hindbrain development, neutrophil differentiation	decreased number of neutrophil, anterior/posterior pattern specification disrupted, decreased hindbrain length	(Li et al 2014b, Xu et al 2009)
28	ENSGACG00000020332	*AR*	male courtship behavior, regulation of transcription	decreased occurrence of male courtship behavior	(Yong et al 2017)
29	ENSGACG00000020333	*MSN / MSNA*	blood vessel lumenization, endoderm development	blood vessel lumenization process quality abnormal, intersegmental vessel unlimenized	(Wang et al 2010)
30	ENSGACG00000010473	*RAB29*	*positive regulation of receptor recycling, transport, mitochondrion organization, Golgi organization (human)*		
31	ENSGACG00000010477	*NPBWR2*	G-protein coupled receptor signaling pathway, neuropeptide signaling pathway		
32	ENSGACG00000009345	*si:dkey-106n21.1*			
33	ENSGACG00000009373	*Kitlg*	melanocyte differentiation	gills and ventrums pigmentation	(Miller et al 2007)
34	ENSGACG00000010758	*si:dkeyp-59c12.1*			
35	ENSGACG00000010762	*GNRHR4*	G-protein coupled receptor signaling pathway		

Note: Zebrafish and human gene ontology (GO) annotations were obtained from the Amigo database (The Gene Ontology 2017)

It is noteworthy that the identification of “strongly adaptive regions” in benthics and limnetics successfully recovered the selective signal in *Kitlg*, known to regulate gill and ventrum pigmentation, which are diverged in Paxton Lake benthics and limnetics (Miller et al 2007) (**Table 3.3, Fig. 3.2**). The adaptive region identified in the analysis lies in the intergenic region upstream of *Kitlg*, which is consistent with the result of previous study showing that divergence in pigmentation is attributed to *cis*-regulatory changes (**Fig. 3.2**). Interestingly, parallel genetic divergence of the intergenic region flanking *Kitlg* was observed in benthics and limnetics from all four lakes (**Fig. 3.2**). Therefore, This intergenic region has diverged in parallel in benthic-limnetic species pairs from the other three lakes*.*

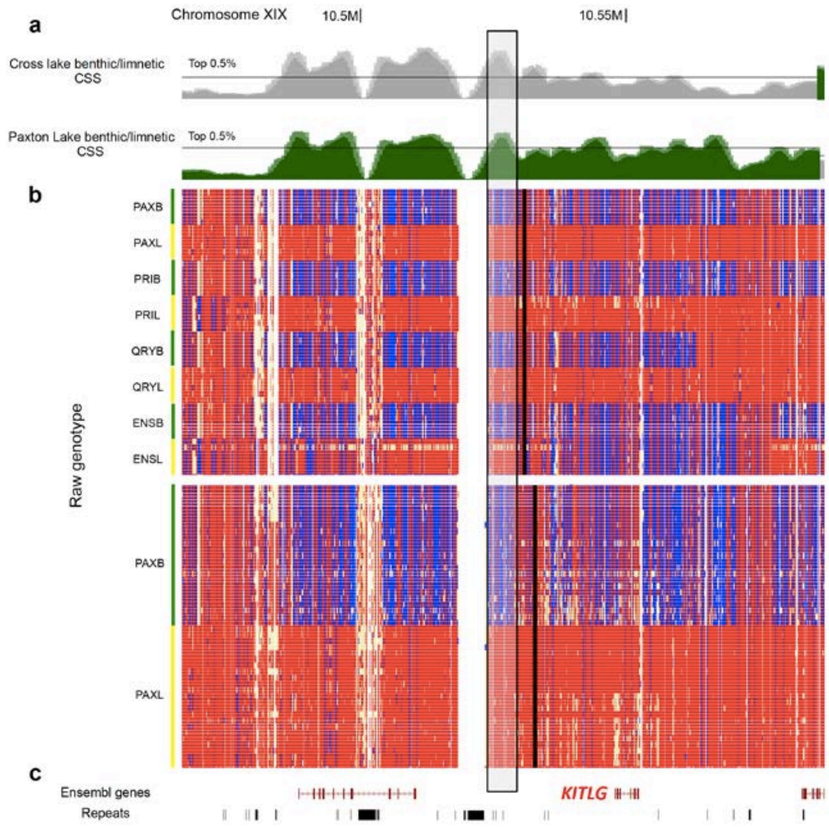

Figure 3.2 | Selective signal at *Kitlg.* a, CSS scores of cross-lake and Paxton Lake benthics and limnetics. **b,** Visual genotype for cross-lake and Paxton Lake benthics and limnetics. Red represents the most frequent allele present in the marine ecotype from Little Campbell River and River Tyne (the ancestral allele), blue the alternative (derived) allele, and yellow the heterozygous allele. **c,** Ensembl gene models and annotated repeat sequences. The vertical shaded box marks the adaptive region identified in the analysis. The white gap in the visual genotype can be attributed to poor alignment of reads to repeat elements.

The lateral line helps fishes to sense peripheral water flow and plays a role in schooling, prey localization, and rheotaxis (Wark et al 2012). It comprises a linear series of punctate specialized hair cells (neuromasts) that run along the lateral midline from anterior to posterior (Ghysen & Dambly-Chaudiere 2004). The density and spatial organization of neuromasts along the lateral line differs between benthics and limnetics (Wark et al 2012). Benthics consistently have more lateral line neuromasts than limnetics, which might be associated with adaptation to divergent light and microhabitat environments (Wark et al 2012, Wark & Peichel 2010). One of the genes (*suppressor of cytokine signaling 3a*, *SOCS3A*) in an adaptive region between benthics and limnetics (chrXI: 9,061,501-9,067,000) lies very close to a QTL marker (chrXI: 9,039,275) associated with the number of neuromasts and lateral plates in benthics (Arnegard et al 2014, Wark et al 2012). *SOCS3* interacts with *signal transducer and activator of transcription 3* (*STAT3*) in a self-restrictive negative feedback loop (Leonard & O'Shea 1998). *STAT3* activates *SOCS3* expression as well as downstream transduction cascades. In turn, *SOCS3* inhibits the expression of its own activator *STAT3*. This self-restrictive feedback loop regulates several biological processes in zebrafish, including cell proliferation, migration, and immune response (Elsaeidi et al 2014, Liang et al 2012, Schebesta et al 2006). Knocking down *SOCS3* or *STAT3* inhibits lateral line neuromast development in zebrafish (Liang et al 2012).

Interestingly, the genetic divergence of *SOCS3* and *STAT3* is different between benthic-limnetic and marine-freshwater species pairs. *SOCS3* is highly diverged in benthics and limnetics as well as in marine and freshwater ecotypes from Little Campbell River but not River Tyne (**Fig. 3.3**). In contrast, *STAT3* is highly diverged between marine and freshwater ecotypes across the Northern Hemisphere (CSS, FDR < 5%)(Jones et al 2012b), but not between benthics and limnetics (**Fig. 3.4**). Thus, although *STAT3* appears to play an important role in marine-freshwater divergence, it does not contribute to the adaptive divergence of benthics and limnetics, as both ecotypes carry the freshwater haplotype. Plausibly, benthic-limnetic divergence of any traits regulated by the *STAT3*/*SOCS3* feedback loop is due to the divergence of *SOCS3* and not to the *STAT3* haplotype common to both. Therefore, it is highly likely that *SOCS3* is the candidate gene for the chromosome XI QTL regulating neuromast development. A detailed analysis of the function of *SOCS3* in benthic and limnetic adaptation can be found in **Chapter 5**.

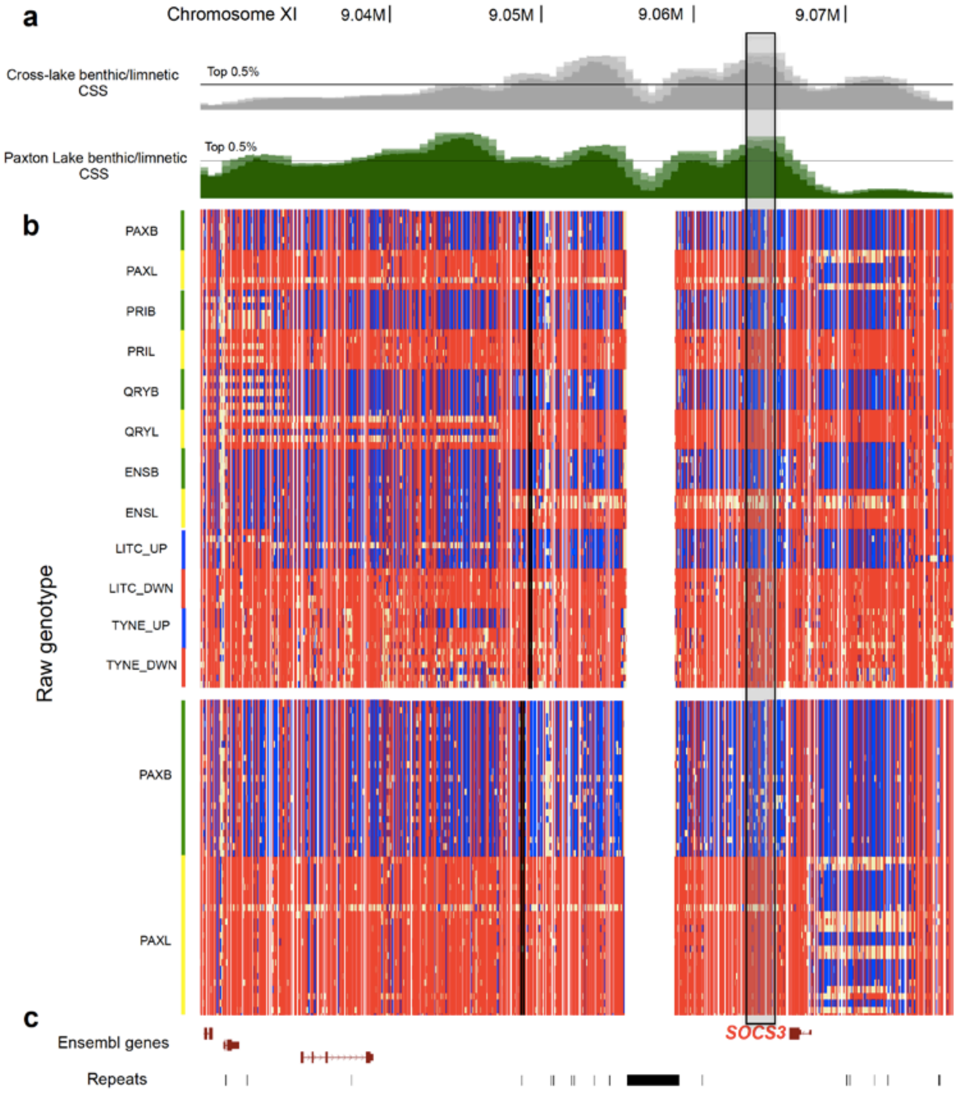

Figure 3.3 | Selective signal at *SOCS3*. a, CSS scores of cross-lake and Paxton Lake benthics and limnetics. Above the horizontal line are the top 0.5% of genome-wide CSS scores. **b,** Visual genotype for cross-lake and Paxton Lake benthics and limnetics as well as marine and freshwater stickleback ecotypes from Little Campbell River (LITC_DWN & LITC_UP) and River Tyne (TYNE_DWN & TYNE_UP). Red represents the most frequent allele present in the marine ecotype from Little Campbell River and River Tyne (the ancestral allele), blue the alternative (derived) allele, and yellow the heterozygous allele. **c,** Ensembl gene models and annotated repeat sequences. The vertical shaded box marks the adaptive region identified in the analysis. The white gap in the visual genotype can be attributed to poor alignment of reads to repeat elements.

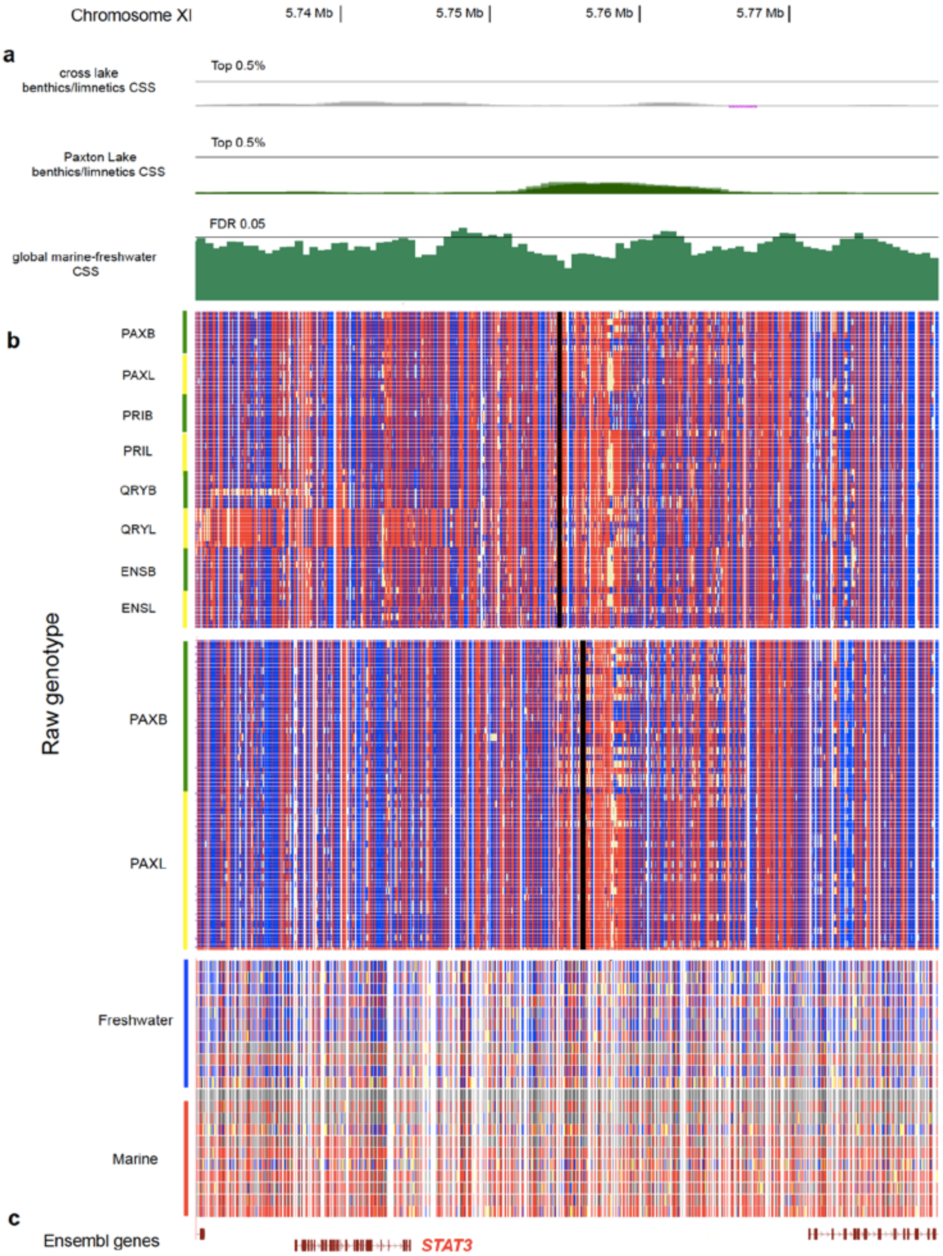

Figure 3.4 | Signature of selection at *STAT3* in benthics and limnetics as well as global marine and freshwater ecotypes. a, CSS scores of cross-lake and Paxton Lake benthics and limnetics as well as global marine and freshwater ecotypes. For CSS scores of global marine and freshwater ecotypes, the horizontal indicates the 5% false discovery rate. **b,** Visual genotype for benthics and limnetics from all four lakes, Paxton Lake benthics and limnetics as well as global marine and freshwater ecotypes. Red represents the most frequent allele in the marine ecotype from Little Campbell River and River Tyne (the ancestral allele), blue the alternative (derived) allele, and yellow the heterozygous allele. **c,** Ensembl gene models. CSS scores and genotypes of global marine and freshwater stickleback ecotypes at *STAT3* region were obtained from (Jones et al 2012a).

3.3.2 Adaptive regions of benthics and limnetics where either benthics or limnetics have been subject to positive selection (“Composite adaptive regions”)

The selection of genomic regions contributing to the adaptation of benthics and limnetics may be incomplete in one or both species or may be difficult to detect using the two approaches applied in this study. To obtain a comprehensive view of the genetic basis of their adaptation, I identified highly divergent genomic regions: those which lie within the top 0.5% of CSS in both Paxton Lake benthics and limnetics and in benthics and limnetics from all four lakes. From these regions, I selected those having either extreme CLR scores (top 0.5%) or at least one SNP of significant nS_L value (FDR < 5%) in either Paxton Lake benthics or limnetics. In the end, 272 genomic regions were recovered as the “composite adaptive regions” of benthics and limnetics.

To characterize the function of adaptive genes, I performed Gene Ontology (GO) enrichment analysis. Genes located in the “composite adaptive regions” and regions 10kb upstream and downstream were used for the analysis. Zebrafish has better syntenic relationship with sticklebacks than other species having GO annotation, while humans have the most extensive GO annotation of any species, and zebrafish of fish species. GO enrichment analysis using human orthologues of stickleback adaptive genes showed significant enrichment of genes involved in ion transmembrane transport, muscle contraction, synaptic assembly, cell-cell signaling, lipid biosynthesis, and collagen fibril organization (**Table 3.4**). GO enrichment analysis using zebrafish orthologues showed significant enrichment of genes involved in lipid transport and in anatomical structure, epithelium, blood vessel, and neural crest morphogenesis (**Table 3.5**).

During breeding, females distinguish conspecific males by their body color and size, which is divergent between benthics and limnetics (Boughman et al 2005). Selection of genes regulating anatomical structure and epithelium morphogenesis may have contributed to this divergent sexual selection. Both human and zebrafish orthologs in “compositive adaptive regions” showed significant enrichment for lipid biosynthesis and transport, In addition, genes involved in fatty acid metabolism are significantly enriched in GO enrichment analysis of adaptive genes of global marine and freshwater sticklebacks (Jones et al 2012a), emphasizing the importance of lipid metabolism during the adaptation of different stickleback populations. This might be related to the different food consumed by benthics and limnetics as well as marine and freshwater sticklebacks.

Interestingly, genes involved in cardiovascular system morphogenesis are significantly enriched using both human and zebrafish orthologues (human GO category: cardiac atrium morphogenesis; zebrafish: blood vessel morphogenesis). This suggests regulation of the development of the cardiovascular system is critical for the adaptation of benthics and limnetics. Temperatures in the benthic zone of a lake are lower than in the littoral zone. Additionally, genes involved in heart development show enrichment in humans (Greenlandic Inuit) and polar bears during adaptation to cold environments (Fumagalli et al 2015, Liu et al 2014). This suggests genes involved in cardiovascular system morphogenesis might contribute to the adaptation of benthics to the colder benthic environment, and to the different oxygen levels in the benthic and limnetic zones of the lakes (Larson 1976).

Table 3.4 Enrichment of Gene Ontology categories in "composite adaptive regions" of benthics and limnetics using human orthologs.

GO category	Annotated	Observed	Expected	P-value	Genes included
ion transmembrane transport	630	24	10.73	0.00017	ITPR1A, CAV3, WWP1, PDE4BA, SLC26A10, GRIK4, ATP13A3, SLC1A5, KCNA10 (1 of many), CACNA2D3 (1 of many), SLC16A1B, ATP8A1, KCNIP2, ARHGEF9B, SLC16A7, SLC41A2B, KCNC2, NLGN2A, CHRNB1,GABRA3, SCN4BB, CHRNE, MINK1, SLC12A9
regulation of muscle contraction	111	8	1.89	0.0006	CAV3, OXTR, PDE4BA, TNNT3A, TNNT2A, JUPA, ADORA2B, SCN4BB
regulation of striated muscle contraction	61	5	1.04	0.00373	CAV3, PDE4BA, TNNT3A, JUPA, SCN4BB
actin-mediated cell contraction	65	6	1.11	0.00081	CAV3, PDE4BA, TNNT3A, TNNT2A, JUPA, SCN4BB
synaptic transmission	592	19	10.08	0.00575	GRM2A, OXTR, ATXN1B, GRIK4, KCNA10 (1 of many), NGFA, STX1A, WNT5A, KCNIP2, ARHGEF9B, SYN3, KCNC2, NLGN2A, CHRNB1, NRXN2 (1 of many), GABRA3, CHRNE, MINK1, GNB2
synapse assembly	69	5	1.18	0.00632	OXTR, WNT5A, NLGN2A, NRXN2 (1 OF many), SPTBN2
sensory perception of pain	59	5	1	0.00323	NR2F6 (1 of many), NGFA, P2RY1, OPRL1, NLGN2A
polyol biosynthetic process	26	3	0.44	0.00947	CRYP27B1, P2RY1, ACER3
membrane lipid biosynthetic process	77	5	1.31	0.00997	CYR61, PIGV, CSNK1G2A, B4GALNT1A, ACER3
protein palmitoylation	25	3	0.43	0.00848	ZDHHC18A, ZDHHC7(1 of many), ZDHHC17
cell-cell signaling	891	25	15.17	0.00885	ITPR1A, GRM2A, OXTR, ATXN1B, GRIK4, KCNA10 (1 OF MANY), NGFA, STX1A, WNT5A, P2RY1, SLC16A1B, JUPA, KCNIP2, ARHGEF9B, SYN3, KCNC2, NLGN2A, FGF11A, CHRNB1, NRXN2 (1 OF MANY), GABRA3, UCP3, CHRNE, MINK1,GNB2
adenylate cyclase-modulating G-protein coupled receptor signaling pathway	104	6	1.77	0.0086	GRM2A, P2RY1, OPRL1, AVPR2 (1 OF MANY), PTGDR2 (1 OF MANY), ADORA2B
single organismal cell-cell adhesion	235	12	4	0.00068	CYR61, COL14A1B, ITGA8, ENSGACG00000009752, WNT5A, JUPA, NLGN2A, NRXN2 (1 OF MANY), MSN (1 OF MANY), PVRL3B, MPZL2B, MINK1
regulation of lymphocyte migration	16	3	0.27	0.00231	WNT5A, SI:DKEY-11F12.2, MSN (1 OF MANY)
collagen fibril organization	37	4	0.63	0.00346	MKXA, COL14A1B, LOXL2A, SERPINH1A
membrane repolarization	22	3	0.37	0.00589	CAV3, KCNIP2, SCN4BB
cardiac atrium morphogenesis	25	3	0.43	0.00848	CYR61, TNNT2A, WNT5A

Table 3.5 Enrichment of Gene Ontology categories in " composite adaptive regions" of benthics and limnetics using zebrafish orthologs.

GO category	Annotated	Observed	Expected	P-value	Genes included
anatomical structure morphogenesis	872	24	14.04	0.0058	PARP3, MCAMB, NAV3, SKIB, HMCN1, TNNT2A, CAV3, SPECC1, PHKG1B, PHACTR4B, TAGLN2, OXTR, FMR1, HDAC3, SOX19A, TTLL3, C1GALT1A, SWAP70B, DNMT3BB.2, CCM2L, JUPA, OPA1, MKXA, MSNB
morphogenesis of an epithelium	206	7	3.32	0.0483	SKIB, TNNT2A, PHACTR4B, SOX19A, SWAP70B, JUPA, MSNB
blood vessel morphogenesis	167	7	2.69	0.018	MCAMB, TNNT2A, PHKG1B, OXTR, HDAC3, C1GALT1A, MSNB
regionalization	175	7	2.82	0.0227	SKIB, USP25, ENSGACG00000020260, FMR1, HDAC3, SOX19A
stem cell differentiation	90	5	1.45	0.0148	PARP3, TNNT2A, PHACTR4B, FMR1, JUPA
organelle organization	362	11	5.83	0.0313	TNNT2A, CAV3, PHACTR4B, TAGLN2, HDAC3, TTLL3, SWAP70B, DNMT3BB.2, JUPA, OPA1, WEE1
neural crest formation	10	2	0.16	0.0106	PARP3, FMR1
lipid transport	51	3	0.82	0.0486	SPNS3, ATP8A1, ENSGACG00000020391
one-carbon metabolic process	22	2	0.35	0.0482	CA4A, CA4B

3.4 Origins of adaptive variation in benthics and limnetics

3.4.1 Benthics and limnetics used standing genetic variation during adaptation

Genetic analyses of adaptive loci demonstrated that both standing genetic variation and *de novo* mutations have contributed to adaptive traits of sticklebacks (Chan et al 2010, Colosimo et al 2005). The “transporter” hypothesis proposed that the adaptive variants segregated in marine populations for a long time before being reused by incipient freshwater populations during rapid adaptation (Colosimo et al 2005, Schluter & Conte 2009). As the sympatric species pairs of sticklebacks can only be found in five out of thousands of lakes in British Colombia (McPhail 1994), benthics and limnetics may have used some *de novo* mutations in their unique adaptation process. To determine the prevalence of adaptive loci originating from the selection of standing genetic variation or *de novo* mutations, I estimated the divergence (coalescence) time of 131 adaptive loci of Paxton Lake benthics and limnetics (**see Section 3.2.2**). The majority of these regions had coalescent times between 75,000 and 200,000 years (**Fig. 3.5**). As ancestral marine sticklebacks started to colonize freshwater habitats as recently as 12,000 years ago, this suggests benthics and limnetics mainly used standing genetic variations already long segregated in stickleback populations during their adaptation.

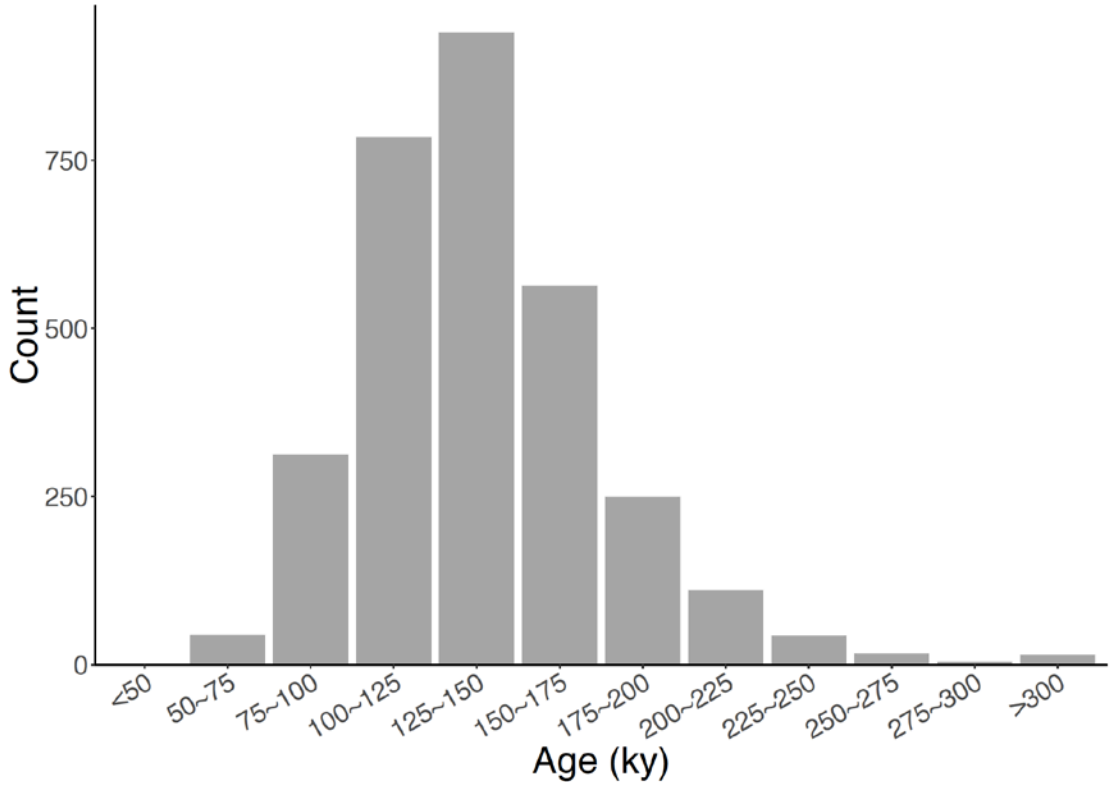

Figure 3.5 | Divergence (coalescence) time of 131 adaptive loci of Paxton Lake benthics and limnetics. Most of the adaptive loci have divergence times older than 100,000 years, suggesting benthics and limnetics mainly used standing genetic variations in their adaptation.

3.4.2 The reuse of genetic variation during adaptation of benthics and limnetics

Benthics and limnetics largely used standing genetic variation during adaptation. In addition, benthics are morphologically and behaviorally similar to freshwater sticklebacks, whereas limnetics possess some morphological and behavioral characteristics similar to marine ecotypes (Rundle & Schluter 2004). This suggests benthics and limnetics might have used genetic divergence similar to that used by marine and freshwater ecotypes. To determine whether benthics and limnetics used the genetic variations mediating marine and freshwater adaptation, I compared genomic pattern of genetic divergence between benthic-limnetic and marine-freshwater ecotype pairs (**Fig 3.6**). The genomic pattern of divergence of benthics and limnetics from all four lakes is not correlated with the pattern of previously described global marine and freshwater ecotypes (Jones et al 2012b) (**Fig 3.6a**). Additionally, most of the adaptive regions of benthics and limnetics (composite adaptive regions) do not have elevated divergence in the global marine-freshwater comparison. Within the "composite adaptive regions" of benthics and limnetics, only 14 (1.1%) showed high genetic divergence between global marine and freshwater ecotypes (CSS, FDR < 5%), indicating that preexisting adaptive alleles which mediated parallel divergence between marine and freshwater ecotypes across the world have contributed very little to adaptive divergence in benthic and limnetic sticklebacks.

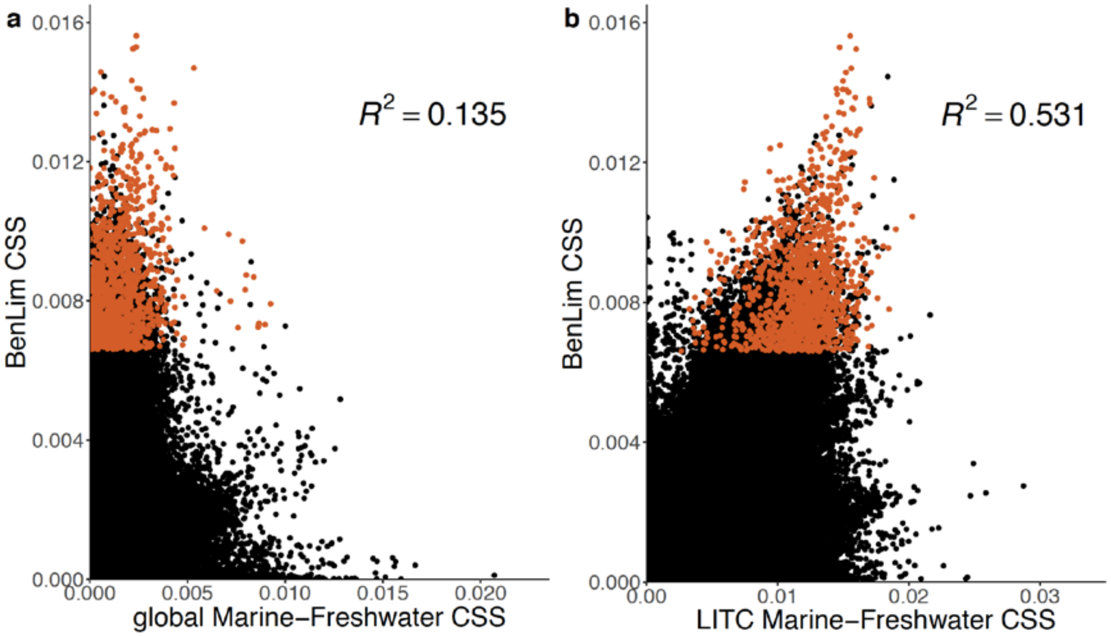

Figure 3.6 | Pairwise comparison of genetic divergence between benthic-limnetic and marine-freshwater stickleback pairs. a, Pairwise comparison of genetic divergence between benthic-limnetic and global marine-freshwater stickleback pairs. CSS scores of previously studied global marine-freshwater (x-axis) (Jones et al 2012b) and cross-lake benthic-limnetic (y-axis) pairs are not correlated (R^2 = 0.135). Most of the divergent regions of benthics and limnetics (orange points; broader set of adaptive regions) are not highly diverged between global marine and freshwater ecotypes. **b,** Pairwise comparison of genetic divergence between benthic-limnetic and LITC marine-freshwater stickleback pairs. CSS scores of LITC marine-freshwater (x-axis) and cross-lake benthics-limnetics pairs (y-axis) are partially correlated (R^2 = 0.531). Many of the divergent regions of benthics and limnetics (orange points; broader set of adaptive regions) are also diverged in LITC marine-freshwater pairs.

As both benthics and limnetics have adapted to a freshwater habitat, they may carry derived (freshwater) haplotypes at adaptive loci of global marine and freshwater ecotypes. To test this hypothesis, I determined the origins of haplotypes (derived or ancestral) at previously identified genomic regions that are consistently divergent between marine and freshwater ecotypes across the Northern Hemisphere (adaptive loci of global marine and freshwater ecotypes, 81 regions) (Jones et al 2012b). More than half (44/76, 57%) of the adaptive loci of global marine and freshwater ecotypes with lengths greater than 350bp have relatively high genetic divergence ($F_{ST} > 0.4$) between marine stickleback ecotypes and both benthics and limnetics from all four lakes, and the divergence between benthics and limnetics is low ($F_{ST} < 0.2$) at these regions (**Appendix Table 13**). This suggests that benthics and limnetics carry similar derived haplotypes at more than half of the adaptive loci of global marine and freshwater ecotypes. This, in turn, suggests that the derived haplotypes at these adaptive loci are critical for the adaptation to freshwater environments. As both benthics and limnetics live in freshwater lakes, the ancestral alleles have no selective advantage at these loci.

The genomic pattern of genetic divergence of benthics and limnetics from all four lakes is correlated with the pattern between a single species-pair of marine and freshwater sticklebacks from the upper and lower reaches of the geographically proximate Little Campbell River (**Fig. 3.6b**). In total, 48.7% of benthic-limnetic “composite adaptive regions” showed high divergence (top 0.5% genome-wide CSS) between marine and freshwater ecotypes from Little Campbell River. This indicates that the adaptive haplotypes underlying benthic-limnetic divergence are also found in geographically proximate populations that do not exist as sympatric benthic-limnetic species pairs.

3.5 Unique genetic divergence of benthics and limnetics

Comparing the genomic pattern of divergence showed that the majority of “composite adaptive regions” in benthics and limnetics have elevated divergence in marine and freshwater ecotypes from Little Campbell River. There are a few “composite adaptive regions” that do not show high genetic divergence between marine and freshwater ecotypes, indicating there may be some genomic regions that are uniquely diverged between benthics and limnetics. Investigating these regions provides valuable insights into their genomic basis and the underlying molecular mechanisms and selective forces driving their adaptive divergence.

To identify the unique divergent regions of benthics and limnetics from all four lakes, I estimated the population-specific genetic divergence of benthics and limnetics with the population branch statistic (PBS) using geographically proximate marine and freshwater ecotypes from Little Campbell River as outgroup populations. PBS quantifies unique allele frequency changes of a population after the point of population split (Yi et al 2010). Several genetic variants, most of them unique to the benthic genome, have high PBS scores. These variants are uniquely fixed in benthics, indicating they are derived alleles (not present in marine sticklebacks) which have been selected in benthics. The bias in benthics over limnetics of alleles with high PBS scores is consistent with the prevalence of selection on derived alleles in benthics (see **Section 2.5.2.2**). Therefore, I focused on genetic variants having extremely high PBS scores in benthics and low PBS scores in limnetics (benthic-specific variants), as they might contribute to the unique adaptation process of benthics. In general, benthic-specific variants are scattered throughout the genome, with only five clusters on three chromosomes (chrIV, chrV, chrXIX) (**Fig. 3.7**). One large cluster of benthic-specific variants was found on the sex chromosome (chrXIX: 19,338,403-19,445,000) (**Fig. 3.7 and 3.8**). The benthic-specific variants in this region have large allele frequency differences ($\Delta p > 0.9$) between benthics and limnetics but no difference ($\Delta p = 0$) between marine and freshwater ecotypes sampled from 5 independent river systems across the Northern Hemisphere, suggesting it diverged only in benthics and limnetics. As there is no gene currently annotated in this region, it possibly contributes to the adaptation of benthics and limnetics as a regulatory element controlling divergent gene expression.

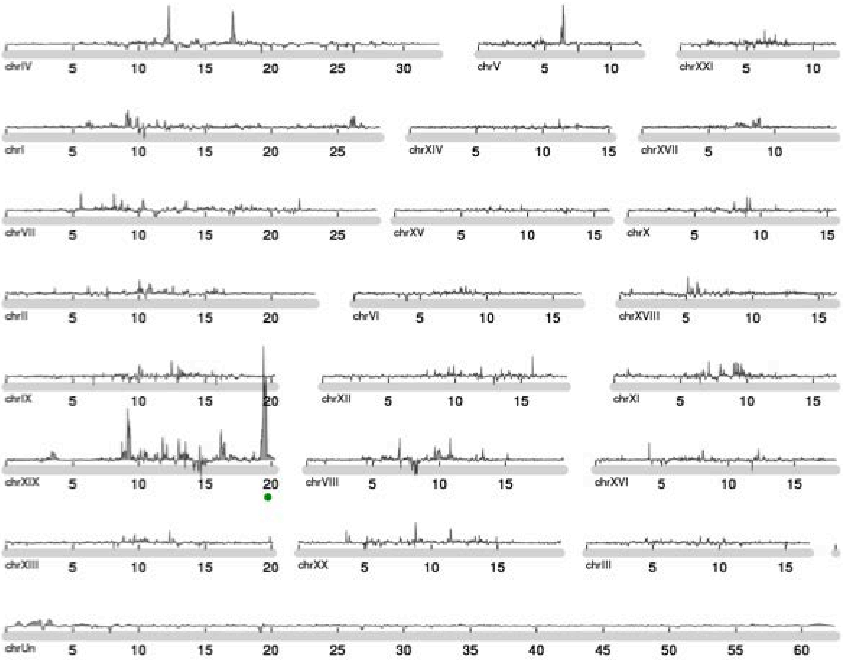

Figure 3.7 | Genomic pattern of population branch statistic (PBS) of benthics and limnetics from all four lakes. Mean PBS values in the sliding windows (size: 1,000bp; step: 200bp) are plotted on the chromosomes. Positive values indicate large PBS in benthics, while negative values indicate large PBS in limnetics. The cluster of benthic-specific variants on chromosome XIX is denoted as a green point below the chromosome.

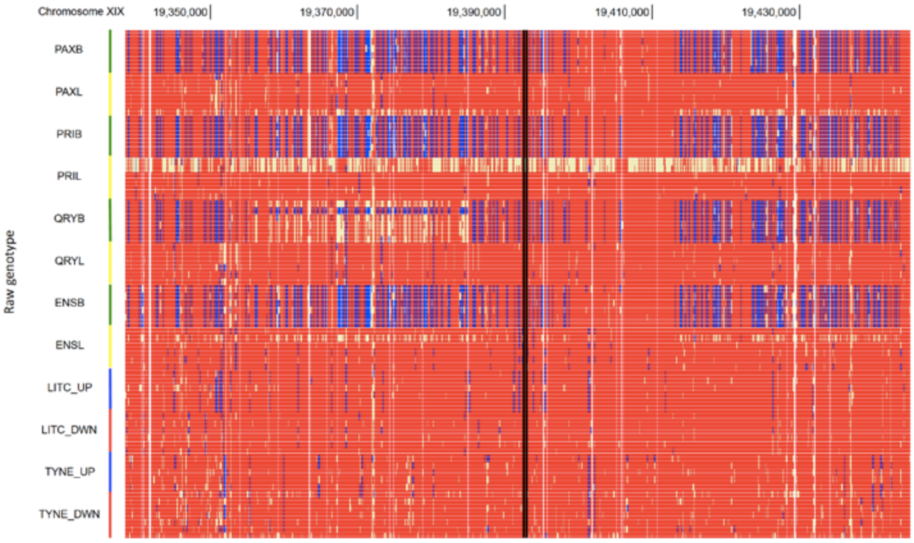

Figure 3.8 | Visual genotype for benthics and limnetics from all four lakes as well as marine and freshwater ecotypes from Little Campbell River (LITC_DWN & LITC_UP) and River Tyne (TYNE_DWN & TYNE_UP) at the cluster of benthic-specific variations on chromosome XIX. Red represents the most frequent allele in marine ecotype from Little Campbell River and River Tyne (the ancestral allele), blue the alternative (derived) allele, and yellow the heterozygous allele.

The genetic variants with high frequency only in benthics contributed to their adaptation after their ancestors colonized the lakes. To find possible genes or regulatory factors contributing to benthic adaptation, I used GO enrichment analysis to characterize the function of genes 1) containing at least two benthic-specific variants in exons or 2) flanking intergenic regions that contained at least five benthic-specific variants. GO enrichment analysis using human orthologs showed significant enrichment in genes involved in cell-cell signaling, organ and kidney morphogenesis, and epithelium, blood vessel, and urogenital and nervous system development (**Table 3.6**). GO enrichment analysis using zebrafish orthologs showed significant enrichment in genes involved in developmental growth, homeostatic processes, inner ear development, and transmembrane transport (**Table 3.7**). The genes containing unique benthic variants were enriched for similar GO categories as the adaptive genes (see **Section 3.3.2**), including transmembrane transport, nervous system development, vascular system development, cell-cell signaling, epithelium development, and anatomical development. These biological processes may be critical to the adaptation of benthics and limnetics. Both *de novo* mutations and standing genetic variation contributed to the divergence of genes involved in these processes.

Table 3.6 Enrichment of Gene Ontology categories for human orthologs of genes containing or flanking genetic variations unique to benthics.

GO category	Annotated	Observed	Expected	*P*-value	Gene included
regulation of nervous system development	474	9	3.23	0.00469	EYA1, EPHB3 (1 of many), NGFRA, SOX9A, SOX8 (1 of many), CIB1, PTPRD (1 of many), NLGN2A
synapse organization	136	5	0.93	0.0023	EPHB3 (1 of many), ADGRL1A, LRRC4.2, NLGN2A, PTRRD (1 of many)
regulation of organ morphogenesis	129	5	0.88	0.00182	EYA1, NGFRA, SOX9A, SOX8 (1 of many), HGF (1 of many)
epithelium development	775	13	5.28	0.00199	EYA1, FEM1B, PRKD2, NGFRA, USH1C, SOX9A, SOX8 (1 of many), HGF (1 of many), TIGARA, PRKX, TRYP1A, TDRD7 (1 of many), RIPK4
epithelium migration	140	5	0.95	0.00261	PRKD2, SOX9A, SOX8 (1 of many), CIB1, ENSGACG00000013796, PRKX
blood vessel morphogenesis	369	8	2.51	0.00348	EYA1, UTS2R, EPHB3 (1 of many), PRKD2, CIB1, ENSGACG00000013796, FGF6 (1of many), PRKX
kidney morphogenesis	46	3	0.31	0.00373	SOX9, SOX8 (1 of many), PRKX
endothelial cell migration	100	4	0.68	0.00474	PRKD2, CIB1, ENSGACG00000013796, PRKX
myoblast differentiation	51	3	0.35	0.00499	SOX9, SOX8 (1 of many), FGF6 (1 of many)
spermatogenesis	249	6	1.7	0.00687	EYA1, FEM1B, EPHB3 (1 of many), SOX9A, SOX8, PRKX
cell-cell signaling	891	14	6.07	0.00243	CACNA1G, EPHB3 (1 of many), KCND2, ENSGACG00000010263, ADGRL1A, SOX9A, SSTR2B, SOX8 (1 of many), LRRC4.2, HGF (1 of many), GRM3, FGF6 (1 of many), EFNB3A, NLGN2A
regulation of fibroblast growth factor receptor signaling pathway	23	3	0.16	0.00049	DUSP6, PRKD2, NGFRA
ERK1 and ERK2 cascade	123	4	0.84	0.00977	DUSP6, PRKD2, SOX9A, CIB1
cellular component movement	1209	17	8.24	0.0026	CACNA1G, ITGA11B, SLC7A10B, TNNI2a.2, ENSGACG00000008376, TNNT3A, EPHB3 (1 of many), PRKD2, NGFRA, ENSGACG00000010263, SOX9, SOX8 (1 of many), CIB1, HGF (1 of many), ENSGACG00000013796, PRKX, EFNB3A
cell adhesion	732	12	4.99	0.00364	ITGA11B, EPHB3 (1 of many), PRKD2, ENSGACG00000010263, ADGRL1A, SOX9, CIB1, ENSGACG00000013796, FGF6 (1 of many), PRKX, PTPRD (1 of many), NLGN2A
positive regulation of behavior	65	3	0.44	0.00979	UTS2R, PRKD2, ENSGACG00000013796

Table 3.7 Enrichment of Gene Ontology categories for zebrafish orthologs of genes containing or flanking genetic variations unique to benthics.

GO category	Annotated	Observed	Expected	*P*-value	Genes included
regulation of developmental growth	23	2	0.13	0.007	DUSP6, SEMA3D
regulation of protein serine/threonine kinase activity	26	2	0.14	0.0089	DUSP6, CCND2A (1 of many)
inner ear development	85	3	0.47	0.0114	EYA1, USH1C, SOX9A
system process	193	4	1.07	0.0215	ENSGACG00000008376, POC1B, USH1C, LIMS2
regulation of biological quality	401	6	2.22	0.0221	CACNA1G, DLDH, SLC4A1A, FLNCA, SEMA3D, ABCB7
homeostatic process	195	4	1.08	0.0222	DLDH, SLC4A1A, FLNCA, ABCB7
organelle assembly	125	3	0.69	0.0315	POC1B, STRA13, TMEM17 (1 of many)
transmembrane transport	438	6	2.42	0.0323	CNCNA1G, SLC7A10B, si:key-106n21.1, KCND2, zgc:77158, ABCB7
cellular component assembly	346	5	1.92	0.0411	POC1B, KCND2, STRA13, TMEM17 (1 of many), NUP93
cell projection organization	252	4	1.4	0.0498	POC1B, USH!C, TMEM17 (1 of many), SEMA3D

3.6 Discussion

3.6.1 Adaptive loci of benthics and limnetics

Benthics and limnetics from different lakes show parallel morphological divergence due to adaptation to contrasting environments. Divergence of several traits between benthics and limnetics has been documented (see **Section 1.6.1**) (McPhail 1994, Schluter & McPhail 1992, Wark & Peichel 2010). However, these traits tend to be obvious because they were mostly quantified by eye.

My study identifying adaptive loci in benthics and limnetics revealed several subtler traits important for adaptation. First, several genes controlling eye development in fish were identified in “strongly adaptive regions”. This suggests benthics and limnetics have divergence in visual ability. Divergence in visual ability has been widely observed in animals (Cuthill et al 2017). Populations sometimes live in environments with different intensities of ambient light (i.e. at different depths of water). In addition, divergent populations of the same species tend to develop different body color patterns for adaptation to local environments and recognition of conspecifics (Cuthill et al 2017). Thus, divergence in visual ability is important for an individual’s adaptation to a local environment and mating preference (Cuthill et al 2017). For example, different ecotypes of African cichlid fish had a wide range of visual sensitivity (Fernald 1984). A female’s preference for conspecific males is based on the male’s body color and thus depends on this variation in visual sensitivity (Fernald 1984, Maan et al 2004). As described in **Section 1.6.2**, benthic and limnetic males gain different nuptial colors during breeding season (McPhail 1994), and females distinguish conspecific males according to nuptial colors (Boughman et al 2005). Therefore, the divergence of benthics and limnetics in visual ability may contribute to their mating preference, and is subject to sexual selection. Second, several genes regulating lipid metabolism and cardiovascular system development were found in “strongly adaptive regions”. Moreover, genes located in “composite adaptive regions” are also enriched for these two processes. This suggests lipid metabolism and cardiovascular system development are important for adaptation of benthics and limnetics to their respective environmental niches. Lipids are one of the most important sources of metabolism in fish (Tocher 2003). A recent study of marine and freshwater sticklebacks showed that freshwater but not marine sticklebacks are exposed to a reduction in nutrient availability during winter. This might due to the temperature decreases in high-latitude freshwater systems during winter, whereas the temperature remains relatively stable in the ocean (Reyes & Baker 2017). Divergence in lipid storage capacity between marine and freshwater sticklebacks may compensate for the difference in food availability (Reyes & Baker 2017). Benthics and limnetics live in different depths – the benthic and littoral zones of a lake, which have different temperatures. Therefore, I hypothesize that the divergence in lipid metabolism ability between benthics and limnetics can be attributed to the differences in food availiability in each zone of the lake during winter. Further study is needed to quantify the divergence of lipid storage between benthics and limnetics, and to investigate the contribution of this divergence to their adaptation. Cardiovascular system development, especially heart development, is crucial for adaptation to cold environments. Recent genomic studies investigating adaptive (selective) regions in Greenlandic Inuit populations and polar bears both identified several genes regulating heart development (Fumagalli et al 2015, Liu et al 2014). Benthics are exposed to lower ambient temperature than limnetics. The adaptation of benthics to a colder environment may explain the high divergence observed in genes controlling cardiovascular system development.

The contributions of several morphological traits to adaptation and speciation of benthics and limnetics have been intensively studied (Schluter 1993, Schluter 1995, Schluter & McPhail 1992). However, the genetic basis of these traits' divergence is largely unknown. Only genes regulating pelvic morphology and gill/ventrum pigmentation have been identified and functionally characterized (Chan et al 2010, Miller et al 2007). My analysis of the adaptive loci of benthics and limnetics identified several important developmental genes that may regulate some adaptive traits in benthics and limnetics, including body size and eye and epithelium development. These genes are candidates for further functional dissection.

3.6.2 The sources of adaptive alleles of benthics and limnetics

Sympatric benthic and limnetic species pairs have only been found in five lakes in British Columbia. A large number of other lakes in British Columbia have just one population of sticklebacks (McPhail 1994, Schluter & McPhail 1992). It is reasonable to hypothesize that benthics and limnetics use unique genetic variation during adaptation and speciation. I demonstrated that the divergence (coalescence) time of “composite adaptive regions” of benthics and limnetics ranges from 75,000 to 200,000 years, which greatly predates the time (~12,000 years ago) when ancestral marine sticklebacks colonized freshwater environments. In addition, I identified a limited number of loci that are uniquely diverged between benthics and limnetics. This suggests benthics and limnetics mainly used standing genetic variation in their adaptation. There is no correlation of patterns of genetic divergence between benthic-limnetic pairs and marine-freshwater ecotype pairs across the Northern Hemisphere, whereas the correlation of patterns of genetic divergence between benthic-limnetic pairs and marine-freshwater ecotype pairs from Little Campbell River is high. This suggests benthics and limnetics largely used pre-existing genetic alleles which mediated marine-freshwater divergence in nearby freshwater systems, but not global marine-freshwater divergence. Benthics and limnetics carry similar derived (freshwater) haplotypes at more than half of the adaptive regions of marine and freshwater sticklebacks across the Northern Hemisphere. These derived alleles are critical for stickleback adaptation to freshwater environments, as both benthics and limnetics reside in freshwater lakes.

Based on these results I hypothesize that the evolution of benthic and limnetic stickleback species pairs largely reused standing genetic variation present in the local geographic region at the time of the double invasion (~4,000 and ~6,000 years ago). The divergent haplotypes of this standing genetic variation were also used by and driven to fixation in nearby freshwater and marine populations, but evolutionary forces unique to the lakes with benthic-limnetic species pairs enabled the maintenance of divergent adaptive haplotypes in freshwater sympatry. Because benthics and limnetics are adapting to freshwater environments, the species pairs can only use a small proportion of the standing genetic variants which mediated global marine-freshwater stickleback divergence, as the derived (freshwater) haplotypes of this standing genetic variation are critical for stickleback's adaptation to freshwater environments.

Investigating the genomic loci (SNPs) that are uniquely diverged between benthics and limnetics provides valuable insight into their recent adaptation to corresponding environmental niches in the lakes, as the genetic alleles specifically fixed in benthics or limnetics were not used in adaptation to other freshwater environments. There are no limnetic-specific alleles and a limited number of benthic-specific alleles in the genomes of limnetics and benthics. Interestingly, genes or regulatory factors containing benthic-specific alleles are enriched for GO categories of epithelium development, cardiovascular system development, and body growth. The "composite adaptive loci" of benthics and limnetics are enriched for genes in these same GO categories, suggesting that these processes are important for adaptation of benthics and limnetics. Selection of standing genetic variants at genes or regulatory factors regulating these processes facilitates rapid adaptation of benthics and limnetics to their corresponding environments, as standing genetic variants are readily available upon a change in environment. Selection of benthic-specific variants at these genes or regulatory factors further increases the fitness of benthics within their environmental niche.

3.7 Materials and Methods

3.7.1 Detailed Analysis of Adaptive Loci of Benthics and Limnetics

3.7.1.1 Identification of adaptive loci of Paxton Lake benthics and limnetics

To identify genomic regions contributed to the adaptation of Paxton Lake benthics and limnetics, I looked for regions highly divergent windows that show strong selective signal in the genome of these two species. As CLR and nS_L have higher accuracy of detecting complete and incomplete selective sweep separately, I combined the selective signals detected by sweepFinder 2 and nS_L together and the adaptive loci were identified as genomic regions having extreme genetic divergence (CSS: top 0.5%) and strong signal of selection detected by sweepFinder 2 (CLR: top 0.5%) or nS_L (nS_L score with FDR < 5%) in both Paxton Lake benthics and limnetics. To identify adaptive loci carrying divergent derived haplotypes in Paxton Lake benthics and limnetics, F_{ST} between marine ecotypes in the dataset (LITC_DWN and TYNE_DWN) and Paxton Lake benthics or limnetics were calculated using VCFtools v0.1.14. Adaptive regions that have high F_{ST} (top 5%) between marine ecotypes and both Paxton Lake benthics and limnetics were selected as carrying divergent derived haplotype in the two species.

3.7.1.2 Identification of adaptive loci of benthics and limnetics

To identify adaptive loci of benthics and limnetics, I looked for adaptive windows (2,500 bp; step: 500 bp) of Paxton Lake benthics and limnetics that are highly diverged between cross-lake benthics and limnetics (CSS: top 0.5%). The overlapping adaptive windows of benthics and limnetics were concatenated into adaptive genomic regions. The genes located in or overlapped with as well as the nearest-neighbor genes on either side of the adaptive regions were identified as adaptive genes of benthics and limnetics.

Visualization of adaptive regions

Paxton Lake and cross-lake benthics and limnetics SNP dataset were uploaded to local UCSC genome browser as custom tracks. The ancestral allele at each SNP was determined according to the most frequent allele of marine ecotypes from Little Campbell River and River Tyne for both dataset, and the derived allele was determined as the alternative allele. Additionally, CSS scores of Paxton Lake and cross-lake benthics and limnetics were uploaded to the genome browser. Ensembl gene build (V68) was used as stickleback gene models for visualization.

3.7.2 Gene ontology enrichment analysis

Gene ontology (GO) enrichment analysis was performed using R package topGO (Bioconductor v2.29.0). Zebrafish and human orthologues of stickleback genes were identified using Ensembl (V90) orthology relationships. As there is no GO annotation for stickleback, I constructed custom GO reference datasets using zebrafish and human genes that have 1-to-1 orthologous relationships of stickleback genes. In total, there are 7,948 zebrafish and 10,570 human genes with GO annotation satisfied with the criteria. The GO hierarchical structure was obtained from the `GO.db` (Bioconductor v3.4.1) annotation and linking zebrafish or human gene identifiers to GO terms was accomplished using `org.Dr.eg.db` (Bioconductor v3.4.1) and `org.Hs.eg.db` (Bioconductor v3.4.1) annotation packages.

GO enrichment analysis for adaptive loci of benthics and limnetics

Genes located within 10kb upstream or downstream of the broader adaptive regions of benthics and limnetics were analyzed for enrichment of GO terms. In total, 289 and 208 genes have 1-to-1 orthologous relationships of zebrafish and human genes separately and their zebrafish or human orthologs were used for GO enrichment analysis. GO categories with *P*-value less than 0.05 and 0.01 for analyses using zebrafish and human orthologs were retained.

GO enrichment analysis for genes containing or flanking benthic-specific variations

Genes containing at least two benthic-specific exon variants or flanking intergenic regions having at least five benthic-specific variants were identified as affected by benthic-specific variants and used in GO enrichment analysis. In total, 85 and 84 of these genes have 1-to-1 orthologous relationships of zebrafish and human genes separately and their zebrafish or human orthologs were used for GO enrichment analysis. GO categories with *P*-value less than 0.05 and 0.01 for analyses using zebrafish and human orthologs were retained.

3.7.3 Comparison of genetic divergence between benthic-limnetic and marine-freshwater stickleback pairs

CSS scores of LITC marine and freshwater ecotypes were calculated in 2,500bp window with 500bp steps using the previously described equation (Jones et al 2012b) with custom Python script. CSS scores of global marine and freshwater ecotypes were downloaded from (Jones et al 2012b). The spearman’s correlation of genetic divergence between benthics/limnetics and global or LITC marine-freshwater stickleback pair was calculated using custom R script. The plots were generated using custom R script.

3.7.4 Divergence time estimation of adaptive loci

Divergence time of Paxton Lake benthics and limnetics adaptive loci was estimated using ARGweaver (Rasmussen et al 2014) v0.8. SHAPEIT phased SNP dataset was converted to ARGweaver input file using custom Python script and input into ARGweaver. Coalescent time was estimated with the following parameters: `-popsize 10,000 --mutrate 6e-8 --recombrate 1.5e-8 -ntimes 40 –maxtime 2e5 –c 10 –n 200`. The mutation rate and recombination rates were estimated using mlRho (Haubold et al 2010) v2.8, which are similar to the estimations in previous study (Roesti et al 2015). ARGweaver partitioned the genome into small intervals that can have the same genealogy and assigned the divergence time to them. The neighboring genomic intervals with the same divergence time estimation were concatenated. And the distribution of divergence time was plotted using custom R script.

3.7.5 Population branch statistics

Population branch statistic (PBS) was calculated for cross lake benthics or limnetics and freshwater ecotypes from Little Campell River (LITC_UP) using marine ecotypes from Little Campell River (LITC_DWN) as outgroup population. I calculated PBS for (benthics, LITC_UP, and LITC_DWN) and (limnetics, LITC_UP, LITC_DWN) triples using the following formula described previously (Huerta-Sanchez et al 2014, Yi et al 2010):

$$\text{PBS} = \frac{T^{A,B} + T^{A,C} - T^{B,C}}{2}$$

, where $T^{A,B}$ = -log(1-$F_{ST}^{A,B}$) is an estimation of the divergence time between benthics and LITC_UP, $T^{A,C}$ is an estimation of the divergence time between benthics and LITC_DWN, and $T^{B,C}$ is an estimation of divergence time between LITC_UP and LITC_DWN. I required that at least 48 alleles (24 individuals) were observed in each population for each SNP used in the F_{ST} calculation. To identify genetic variations unique to benthics, I subtracted PBS of limnetics from PBS of benthics and kept top 0.1% of the results as candidate variations.

4 EVOLUTIONARY HISTORY OF BENTHICS AND LIMNETICS

4.1 Background and Aims

The patterns of genomic diversity within and between populations are not only shaped by natural selection but also the demographic history of the population (Ellegren 2014). Genetic variations can be fixed and removed from the population due to historical population bottlenecks and expansions (Hedrick 2005). In addition, gene flow and introgression can reduce the genetic divergence between two populations (Sousa & Hey 2013). Therefore, it is critical to determine the demographic history in the study of adaptation of a species.

Determining the prevalence of sympatric and allopatric speciation in nature is one of the important and controversial subjects of evolutionary biology (Coyne & Orr 2004). Sympatric speciation was considered as uncommon due to the famous critiques of Mayr and scarce of examples in nature (Coyne & Orr 2004). However, with the advance in theoretical studies of speciation and advent of genomic era, sympatric speciation has been shown to be possible (Bolnick & Fitzpatrick 2007).

Sympatric benthics and limnetics are considered to evolve from sympatric speciation because of the discovery of character displacement and disruptive sexual selection in the species pair (Boughman et al 2005, Rundle & Schluter 2004, Schluter & McPhail 1992). In contrast, recent studies of evolution of benthics and limnetics supported allopatric with double-invasion hypothesis (see **Section 1.6.2**) (Jones et al 2012a, Taylor & McPhail 2000). The double-invasion hypothesis predicts several properties of benthics and limnetics: 1) species pair from the same lake should have a polyphyletic origin; 2) assuming similar effective population sizes on colonization, the benthics would have less genetic diversity than limnetics as drift and selection have had more time to fix variations in benthics; 3) limnetics should be genetically closer to marine sticklebacks than benthics; Previous phylogenetic study of benthics and limnetics using six microsatellite identified polyphyletic origin of species-pair in the same lake, which is consistent with the prediction of allopatric speciation (Taylor & McPhail 2000). However, However, two of the phylogenetic trees generated in the study were ambiguous due to limited number of markers. A recent genomic study using makers generated by SNP genotyping array identified two features that are consistent with the prediction of double-invasion hypothesis: 1) lower genetic diversity of benthics compared to limnetics, 2) closer genetic relationship of marine sticklebacks with limnetics than benthics. Nevertheless, less heterozygosity of benthics and closer relationship of limnetics and marine sticklebacks can arise from benthics and limnetics experiencing different effective population size changes. Finally, it is also possible and even likely that both the double-invasion hypothesis of allopatric divergence and the sympatric speciation hypothesis are correct: these species pairs may have evolved as a result of initial divergence in allopatry followed by secondary contact via double invasion and be subject to ongoing divergent selection pressures in sympatry that drive character displacement and divergent sexual selection. I aim to shed more light on the evolution of these two species by resolving their ancestry and determining their demographic history using high-density genetic markers.

In this chapter, I study the evolutionary history of benthics and limnetics from different aspects:

- I determine the best-fit demographic model of benthics and limnetic speciation.
- I investigate the history of population size change of benthics and limnetics
- I identify the populations that share most ancestry with benthics and limnetics.

4.2 The ancestry of benthics and limnetics

4.2.1 Genetic relationship of benthics and limnetics as well as marine and freshwater sticklebacks

To identify the ancestry of benthics and limnetics, I first resolved the phylogenetic relationship of benthics and limnetics in the context of a global set of marine and freshwater sticklebacks. To determine the genetic relationship of benthics and limnetics, I first performed phylogenetic analysis of benthics and limnetics from all four lakes as well as 210 individuals of marine and freshwater ecotypes. A maximum likelihood (ML) tree was constructed for benthics and limnetics from all four lakes as well as 210 marine and freshwater sticklebacks across the Northern Hemisphere. The ML tree was constructed using genome-wide autosomal SNPs with minor allele frequency (MAF) greater than 0.01. The freshwater individual collected in Gifu, Japan (GIFU) was used as a outgroup (**Fig. 4.1**). Stickleback individuals collected along the Pacific and Atlantic Ocean formed two distinct clades. Within the Pacific and Atlantic clades, marine and freshwater ecotypes formed distinct clades. Atlantic and Pacific marine sticklebacks are close to the root of the tree, indicating closer genetic relationship between marine sticklebacks and GIFU. Pacific freshwater sticklebacks formed three distinct clades (California, Alaska, and British Columbia) according to their geographic origins. In general, freshwater ecotypes have longer branch length than marine ecotypes. This may due to lack of gene flow between freshwater populations (unlike marine “panmixia”), and that adaptation to freshwater environment involves strong bottlenecks and rapid fixation of a subset of standing genetic variation.

Limnetics from three lakes (except Enos Lake) formed a monophyletic clade and do not cluster with other marine or freshwater populations. Benthics from all four lakes cluster with freshwater ecotypes from the geographically proximate Little Campbell River (LITC_UP), suggesting benthics are genetically close to this freshwater population. In addition, Enos Lake limnetics cluster with Enos Lake benthics, suggesting strong directional gene flow from Enos Lake benthics to limnetics. The clustering of benthics and limnetics by species but not by lakes suggests the species pair do not derive from a single ancestral population within each lake, and strengthens the evidence of allopatric speciation. Similar to marine ecotypes, the branch lengths of limnetics are shorter than the lengths of benthics. This indicates benthics are more diverged from ancestral marine population than limnetics, which is consistent with the prediction of double-invasion hypothesis.

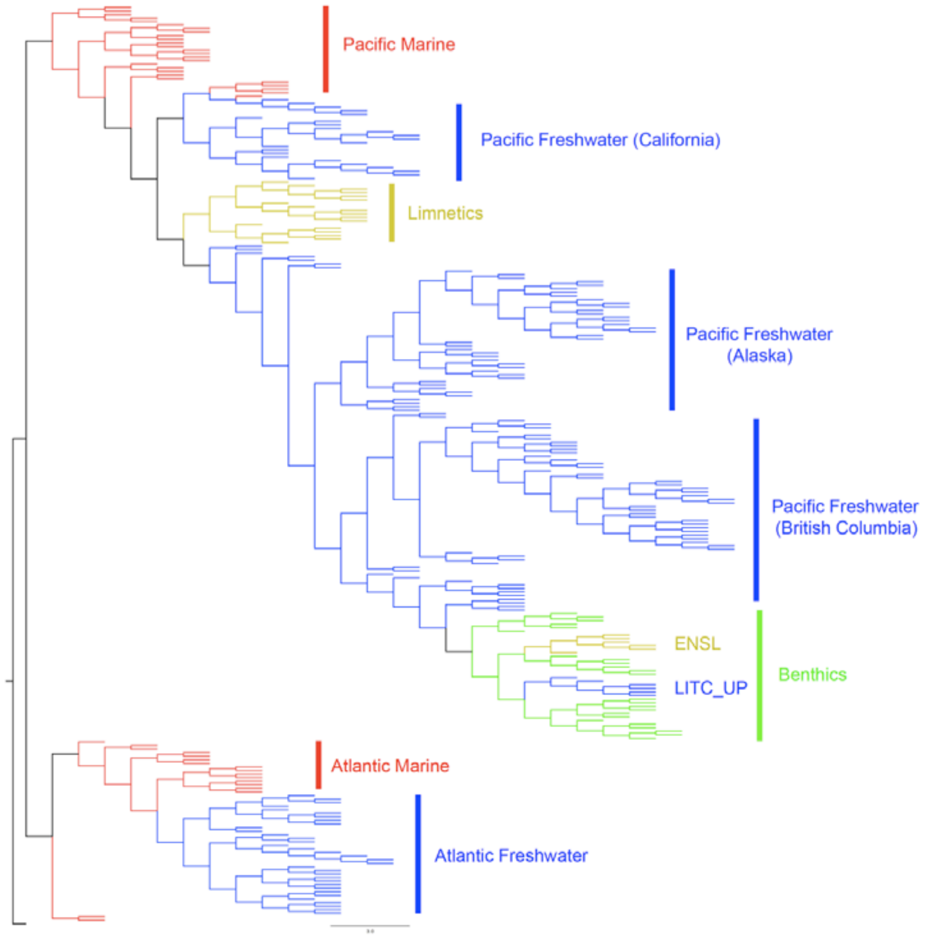
Pacific Marine
Pacific Freshwater (California)
Limnetics
Pacific Freshwater (Alaska)
Pacific Freshwater (British Columbia)
ENSL
LITC_UP
Benthics
Atlantic Marine
Atlantic Freshwater

Figure 4.1 | Maximum likelihood (ML) tree of benthics and limnetics from all four lakes as well as 210 marine and freshwater sticklebacks. Benthics cluster with freshwater individuals from nearby Little Campbell River (LITC_UP), while limnetics form a monophyletic group and do not cluster with other marine and freshwater populations.

Although constructing a phylogenetic tree is a common method to infer genetic relationships among populations, a bifurcating tree simplifies these relationships by considering only population splits without gene flow and assumes that the ancestral alleles are not present in the modern day sample (Pickrell & Pritchard 2012). To overcome this problem, the TreeMix program estimates a maximum likelihood tree of a set of populations using their allele frequency given a Gaussian approximation and builds a residual matrix of fits of populations to the initial tree (Pickrell & Pritchard 2012). The positive residuals indicate a closer relationship between populations than as shown in tree, while negative residuals indicate a more distant relationship. Migration and gene flow events would then add to populations that have poor fits in the residual matrix. I used TreeMix program to infer the genetic relationship of benthics and limnetics sticklebacks as well as marine and freshwater populations with 5 or more individuals from the larger 210 genome dataset. The ML tree of benthics and limnetics as well as marine and freshwater populations was first constructed by TreeMix program using genome-wide SNPs with marine population from Big River, California (BIGR_DWN) as a outgroup (**Fig. 4.2a**) because the previously used outgroup described above (GIFU) is a singleton individual. Similar to the conventional ML tree, Pacific and Atlantic stickleback populations showed large divergence and formed distinct clades.

Benthics cluster with freshwater population from Little Campbell River (LITC_UP) and Bonsall Creek in British Columbia (BNST). Limnetics do not cluster with other marine or freshwater populations. Benthics have larger estimated drift coefficient (longer branch length) than limnetics, suggesting benthics derived from ancestral population earlier than limnetics, and drift had more time to fix/remove variation in the genome of benthics than limnetics. Enos Lake limnetics cluster with Enos Lake benthics. Furthermore, the comparison of benthics and limnetics from Enos Lake has the largest positive residual (**Fig. 4.2b**). It indicates Enos Lake benthics and limnetics have closer genetic relationship than species pairs from other lakes, which is consistent with the increased gene flow between these two species. The likelihood of TreeMix ML tree substantially improved after adding migration events (**Fig. 4.2a**). TreeMix identified gene flow from Paxton Lake benthics to Paxton Lake limnetics and mutual gene flow between benthics and limnetics from Little Quarry Lake when 3 migration events added to the tree. This suggests the gene flow is higher between species pairs from Paxton and Little Quarry Lakes than from Priest Lake.

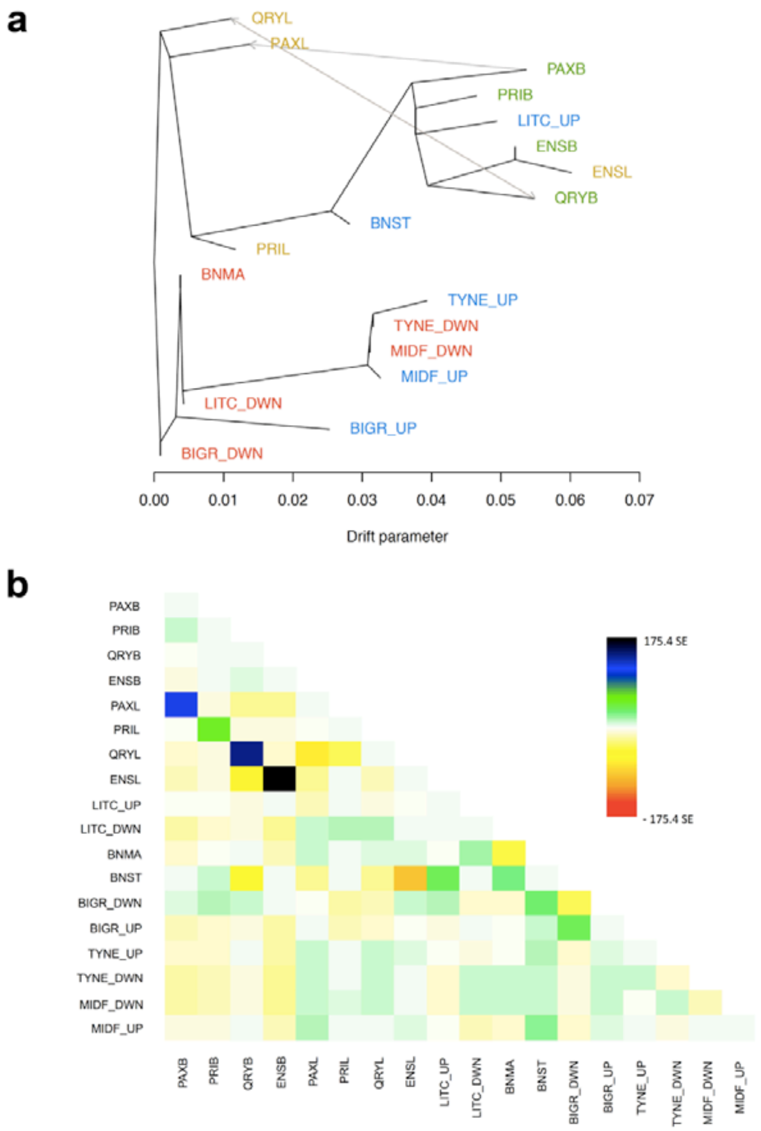

Figure 4.2 | Genetic relationship of benthics and limnetics as well as hybrid zone marine and freshwater populations identified by TreeMix. a, Maximum likelihood (ML) tree of benthics/limnetics and hybrid zone marine/freshwater populations based on allele frequency. Three migration events were added and shown as grey arrows. **b,** Matrix of residues from the fit of data to the data. The positive residues indicate closer relationship between populations than as shown in tree, while negative residues indicate distant relationship. Refer **Table 2.1** for population code of marine and freshwater populations. PAXB: Paxton Lake benthics; PAXL: Paxton Lake limnetics; PRIB: Priest Lake benthics; PRIL: Priest Lake limnetics; QRYB: Little Quarry Lake benthics; QRYL: Little Quarry Lake limnetics; ENSB: Enos Lake benthics; ENSL: Enos Lake limnetics.

To determine the genetic relationship of benthics and limnetics, I performed PCA of benthics and limnetics in the context of global marine and freshwater sticklebacks using three variant datasets (all variants, neutral variants, and variants under selection). PCA of benthics and limnetics was first performed using genome-wide SNPs, which is described previously in **Section 2.3.1**. When projected onto the PC space of benthic and limnetic sticklebacks, marine and freshwater individuals were only separated by the first principal component (PC1), where marine and freshwater populations are placed close to limnetics and benthics respectively (**Fig. 4.3a**). This suggests the genomic divergence between benthics and limnetics resembles the divergence between marine and freshwater sticklebacks. Similar to the result of phylogenetic reconstruction, freshwater sticklebacks from the Little Campbell River is placed closer to benthics than other marine or freshwater populations, while PCA places no population close to limnetics. Although the second principal component (PC2) separates benthics and limnetics by lakes, marine and freshwater ecotypes do not separate on PC2, indicating benthics and limnetics from different lakes have unique genetic variation that does not segregate among marine or freshwater populations in the broader 210 genome dataset. These variations might arise from the adaptation of benthics and limnetics to the unique environment of each lake, which is consistent with the prediction of parallel evolution of benthics and limnetics (Rundle et al 2000, Taylor & McPhail 1999).

As the analyses described previously showed several genomic regions of benthics and limnetics have been subject to natural selection (see **Section 2.4.3**), I performed PCA using SNPs from "parallel non-divergent regions" of benthics and limnetics (see **Section 2.4.2**) to remove the effect of selection. PCA using neutral variants showed a distinct result from PCA using genome-wide SNPs (**Fig. 4.3b**). Benthics and limnetics from the same lake cluster together in the analysis. PC1 and PC2 explain similar amount of variation in the genome (9.6% vs. 9.1%). PC1 and PC2 both separate benthics and limnetics by lakes. This suggests benthics and limnetics from the same lake have close genetic relationship in neutral genomic regions, which may be derived from the gene flow in neutral regions. Interestingly, marine and freshwater sticklebacks do not separate and formed a single cluster when projected onto the benthics and limnetics PC space. It indicates that there is unique genetic variation in benthics and limnetics at neutral genomic regions. As this unique variation does not contribute to the divergence of benthics and limnetics as well as marine and freshwater sticklebacks, they might evolve from the unique demographic history of benthics and limnetics.

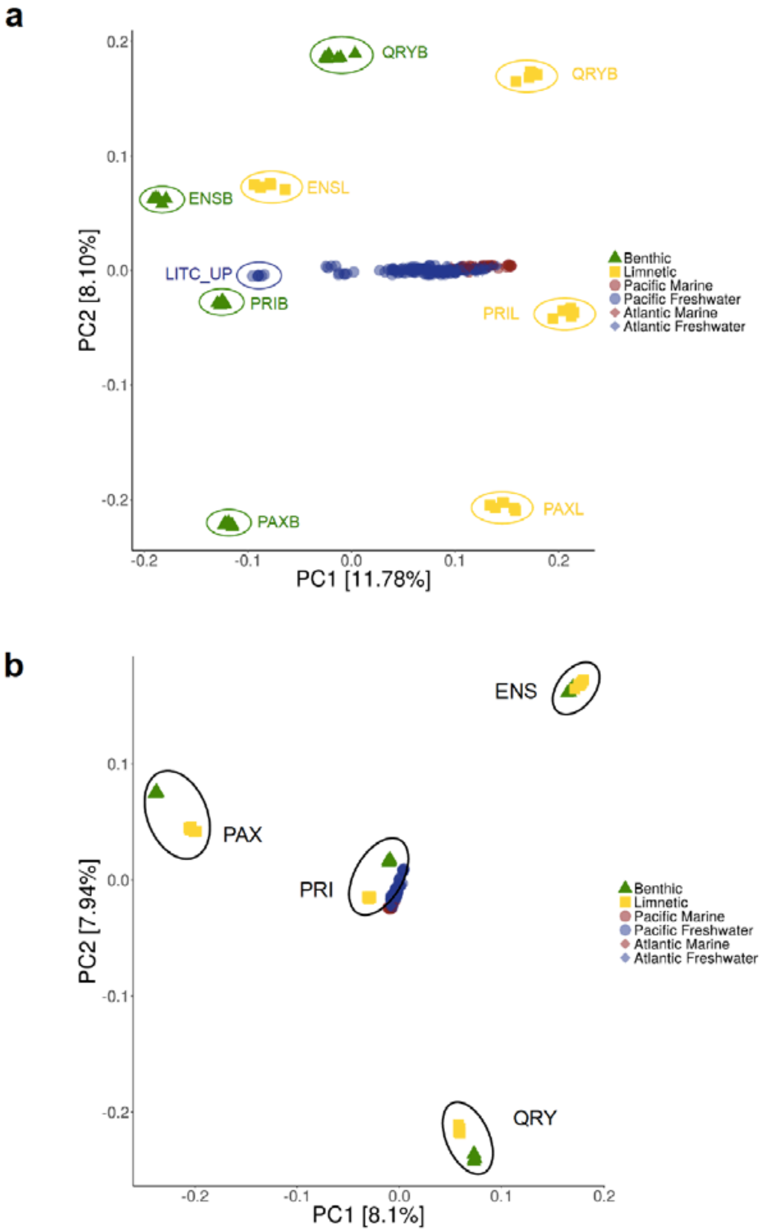

Fig. 4.3 | Principal component analysis (PCA) of benthics/limnetics and a global set of marine and freshwater sticklebacks. PCA was first performed for benthics and limnetics from all four lakes, and then marine and freshwater sticklebacks were projected onto the PC variation space of benthics and limnetics. **a,** PCA of benthics/limnetics and marine/freshwater sticklebacks using genome-wide SNPs. Freshwater ecotypes from Little Campbell River (LITC_UP) show a close genetic relationship with benthics. **b,** PCA of benthics/limnetics and marine/freshwater sticklebacks using neutral SNPs. Benthics and limnetics are separated by lake. Marine and freshwater sticklebacks do not separate in the analysis.

Freshwater ecotypes from Little Campbell River show a close genetic relationship with benthics in both PCA and phylogenetic analysis, which suggests this freshwater population from a geographically proximate river may share most ancestry with benthics. To further investigate the ancestry of benthics and limnetics, I calculated outgroup f_3 statistics for benthics/limnetics and marine/freshwater populations (Patterson et al 2012). Outgroup f_3 statistic has been widely used in population genetic analyses to investigate patterns of admixture and shared ancestry of a population (Pickrell & Reich 2014, Sousa & Hey 2013). The statistic evaluates shared drift between two populations from a common outgroup (which is highly diverged from test populations) by measuring allele frequency correlations between populations. More shared drift between two populations implies they share more ancestry with each other. Larger outgroup f_3 scores indicate more shared ancestry between two populations. I calculated outgroup f_3 between benthic/limnetic and Pacific marine/freshwater populations with more than 4 individuals in the SNP dataset of benthics, limnetics, and global marine/freshwater sticklebacks (see **Section 2.2**). As there is large genetic divergence between Pacific and Atlantic stickleback populations, marine population from River Tyne (TYNE_DWN) was used as the outgroup (**Fig. 4.4**). Freshwater populations from Little Campbell River (LITC_UP) and Bonsall Creek (BNST) populations have substantially higher outgroup f_3 scores with benthics from all four lakes than other marine or freshwater populations, indicating these two populations shared most ancestry with benthics. In contrast, no clear population with shared ancestry is identified for limnetics from three lakes (except Enos Lake). Enos Lake limnetics has notably larger outgroup f_3 scores with freshwater populations from Little Campbell River and Bonsall Creek stickleback populations than other marine or freshwater populations, suggesting these two freshwater populations shared more ancestry with Enos Lake limnetics than other marine or freshwater populations. This may arise from the increased gene flow from Enos Lake benthics to limnetics.

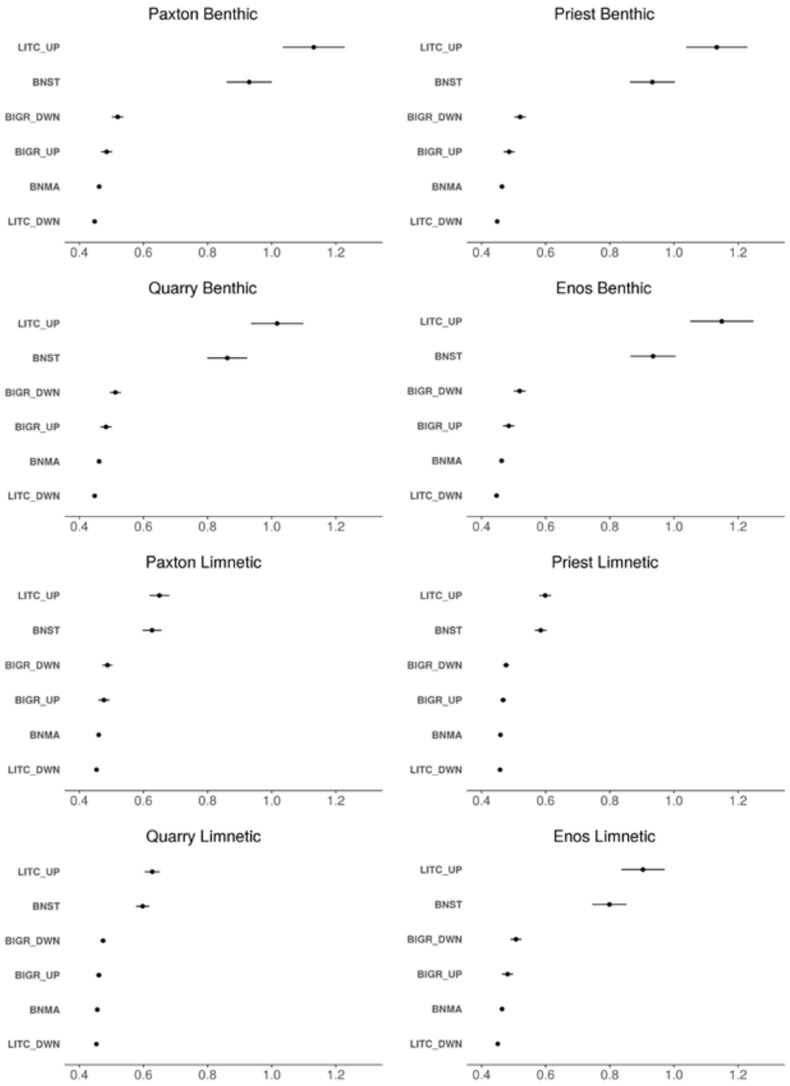

Figure 4.4 | Outgroup *f_3* scores between benthics/limnetics and Pacific marine/freshwater stickleback populations. The standard errors were estimated using jackknife resampling and indicated as bars. LITC_UP: freshwater ecotypes from Little Campbell River, BNST: freshwater ecotypes from Bonsall Creek, BIGR_UP: freshwater population from Big River, California, BIGR_DWN: marine population from Big River, California, BNMA: marine population from Bonsall Creek, LITC_DWN: marine population from Little Campbell River.

4.3 Demographic history of Paxton Lake benthics and limnetics

4.3.1 Population size history of Paxton Lake benthics and limnetics

Previous studies of benthic and limnetic evolution found indirect evidence supporting the double-invasion hypothesis (Jones et al 2012a, Taylor & McPhail 1999). However, the detailed demographic model of benthics and limnetics speciation is still lacking. Several algorithms/programs have been developed to infer the demographic history of populations using dense SNP markers generated from whole genome resequencing studies (Schraiber & Akey 2015). Therefore, I tried to infer the demographic model of Paxton Lake benthics and limnetics using two approaches. The density of heterozygous variants is higher in genomic regions with long coalescence time (time to the most recent common ancestor) than regions with short coalescence time, and the density of heterozygous variants varies along the chromosome due to recombination. Thus, the local density of heterozygous variants can be used to infer the local coalescence time across the genome. The SMC++ program infers the historical population size of test population by evaluating the distribution of coalescence time for alleles from a large set (up to hundreds) of individuals.

The histories of ancestral population size of Paxton Lake benthics and limnetics from all four lakes were inferred using SMC++ with 23 Paxton Lake benthics and 23 Paxton Lake limnetics respectively. To remove the effect of natural selection, only SNPs on the putatively “neutral” chromosome (chrXV) were used in the analysis. The ancestral population size was inferred by assuming a mutation rate of 6x10^{-8}, which is used by previous study and estimated using the SNPs from the input dataset of SMC++ analysis (Roesti et al 2015). Both Paxton Lake benthics and limnetics have experienced a decline of population size between 20,000~30,000 years ago followed by an expansion of population size (**Fig. 4.5**). As both Paxton Lake benthics and limnetics experienced the decline of population size at similar time interval, this may result from a split of ancestral marine population. Starting around 9,000 years ago, Paxton Lake benthics and limnetics experienced a decline of population size followed by population size expansion. The decline started about 2,000 years earlier in Paxton Lake benthics (about 7,000 years ago) than limnetics (about 5,000 years ago), which may correspond to the different time when the ancestors of these two species colonized freshwater habitats or Paxton Lake. This is consistent with the prediction of double-invasion hypothesis. The reduction of population size in Paxton Lake benthics (smallest population size: ~1,000) is two times more severe than in Paxton Lake limnetics (smallest population size: ~2,000). In addition, the population size expansion occurred about 500 years earlier in Paxton Lake limnetics than in Paxton Lake limnetics. This may result from the stronger natural selection in Paxton Lake benthics.

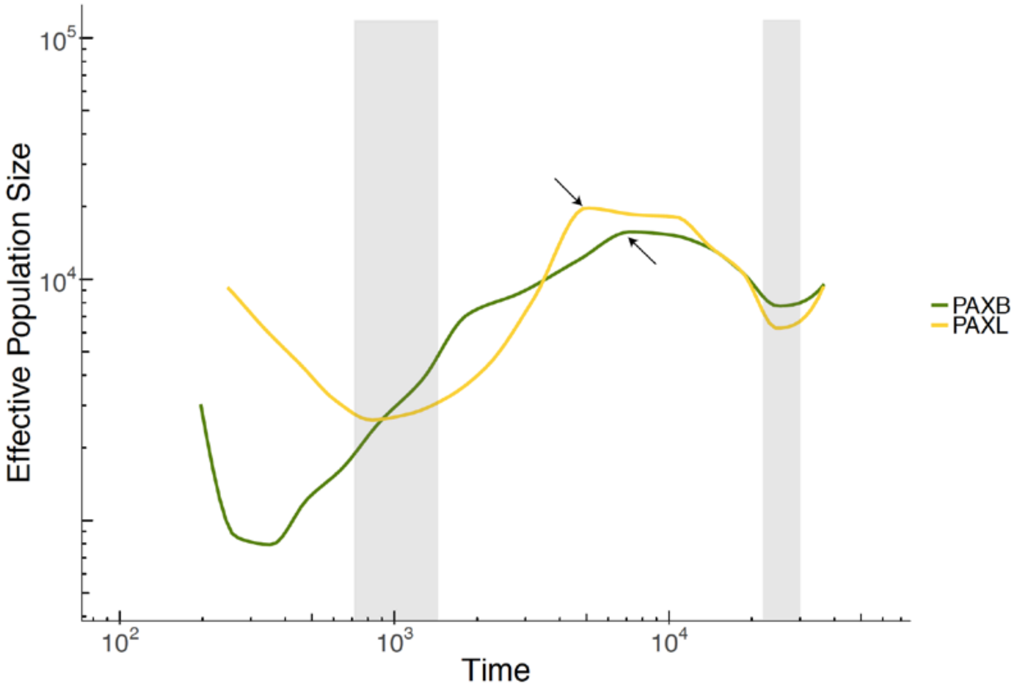

Figure 4.5 | Inferred historical population size of Paxton Lake benthics and limnetics. Time in history was estimated by assuming a generation time of 1 year and a mutation rate of 1.5x10^{-8}. The historical population bottlenecks of Paxton Lake benthics and limnetics are indicated by shared rectangles. The starts of recent population size decline in Paxton Lake benthics and limnetics are indicated by arrows on the plot.

4.3.2 Demographic model of Paxton Lake benthics and limnetics inferred by δaδi program

Demographic inference by estimating historical population size is useful and important, whereas gene flow between populations is another important factor that shapes the genomic pattern of genetic variation. However, SMC++ cannot infer the gene flow between populations. Thus, to comprehensively investigate the joint demographic history of Paxton Lake benthics and limnetics, I infer the demographic model of them using the δaδi program (Gutenkunst et al 2009b). δaδi can infer the demographic model of up to three populations by fitting a simulated joint allele frequency spectrum (two-dimensional or three-dimensional) to joint allele frequency spectrum that is empirically observed. It can be used to identify the best demographic model of test populations according to the fit of the simulated joint allele frequency spectrum to the empirical spectrum. In addition, the program infers divergence time, migration rate, and population size history of test populations in a given model.

I used the δaδi program to infer the joint demographic history of benthics and limnetics using 23 Paxton Lake benthics and 23 Paxton Lake limnetics. As the δaδi program assumes the underlying variants used in the analysis are selectively neutral, only SNPs in the genomic regions that are not diverged (CSS score, *P*-value > 0.5) between Paxton Lake benthics and limnetics (see **Section 2.5.1**) were used for the analysis. The ancestral allele at each SNP was determined as the most frequent allele in marine sticklebacks from Little Campbell River and River Tyne. A total of 2,667,791 SNPs were used to construct the two-dimensional unfolded allele frequency spectrum. Three demographic models of allopatric speciation (Allopatric-1, Allopatric-2, Allopatric-3) and one model of sympatric speciation were tested with different settings for migration rate and population size changes (**Fig. 4.6**). As the demographic inference using SMC++ revealed the recent decline of population size started earlier in Paxton Lake benthics than in limnetics, all three tested demographic models of allopatric speciation have Paxton Lake benthics diverged from ancestral population earlier than limnetics. All demographic model of allopatric speciation have higher Poisson likelihoods than the model of sympatric speciation in the fitness test, indicating Paxton Lake benthics and limnetics are unlikely to evolve from sympatric speciation.

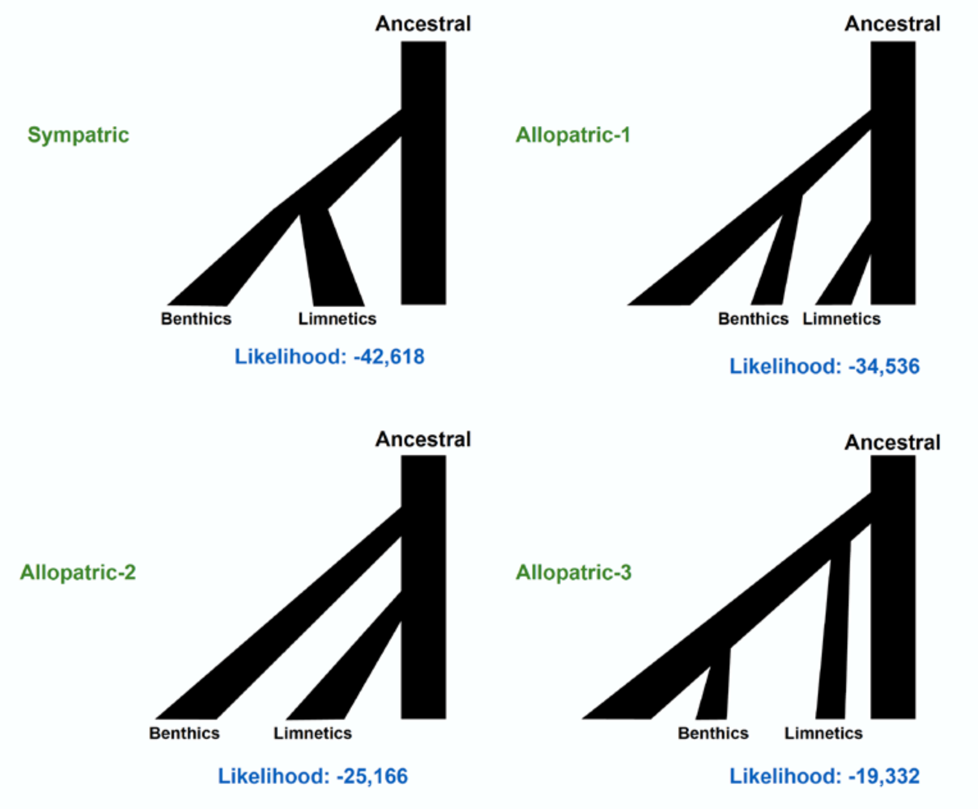

Figure 4.6 | The likelihoods of four demographic models of Paxton Lake benthics and limnetics in demographic inference using δaδi. Three demographic models of allopatric speciation (allopatric-1, allopatric-2, allopatric-3) and one sympatric model were tested. The model of sympatric speciation has lower likelihood than all allopatric models, suggesting Paxton Lake benthics and limnetics were not derived from sympatric speciation. One of the models of allopatric speciation (allopatric-3) has the highest likelihood and was used in subsequent analysis.

I identify the maximum-likelihood model parameters of the best-fit demographic model of Paxton Lake benthics and limnetics (Allopatric-3) using non-linear optimization. The δaδi program assumes all the input SNPs are independent (not-linked) to each other. However, SNPs in my dataset is not completely independent. Therefore, to remove the effect of linkage disequilibrium, I determined the confidence intervals of each model parameter using conventional bootstraps. In total, maximum-likelihood model parameters were estimated for 100 bootstrap datasets, and 95% confidence intervals (95% C.I.) were determined (**Fig. 4.7**). In allopatric-3 model, the ancestral population of benthics and limnetics diverged from the main ancestral population between 26,840 to 30,006 years ago (95% C.I.). Then the ancestral population of benthics (95% C.I.: 25,875~28,764 years ago) and the ancestral population of limnetics (95% C.I.: 996~1,169 years ago) diverged from the common ancestral population separately (**Fig. 4.7b**). There is bidirectional gene flow between Paxton Lake benthics and limnetics, with the gene flow from benthics to limnetics (95% C.I.: $2.87x10^{-3}$~$3.17x10^{-3}$, migration rate) substantially higher than from limnetics to benthics (95% C.I.: $3.69x10^{-4}$~$4.29x10^{-4}$, migration rate) (**Fig. 4.7a**). This indicates Paxton Lake benthics and limnetics diverged from ancestral population at different time in the history, and there are gene flows between Paxton Lake benthics and limnetics after they cohabited in the same lake, which is consistent with the allopatric speciation following secondary contact model. The gene flow from Paxton Lake benthics to limnetics is 10 times higher than from limnetics to benthics, which may due to the introgression of freshwater adaptive alleles from benthics to limnetics. Consistent with the estimation based on the genomic heterozygosity and linkage disequilibrium (see **Section 2.3.1**), the recent population size of Paxton Lake benthics (95% C.I.: 1,959~2,157) is smaller than limnetics (95% C.I.: 3,442~3,798) (**Fig. 4.7b**).

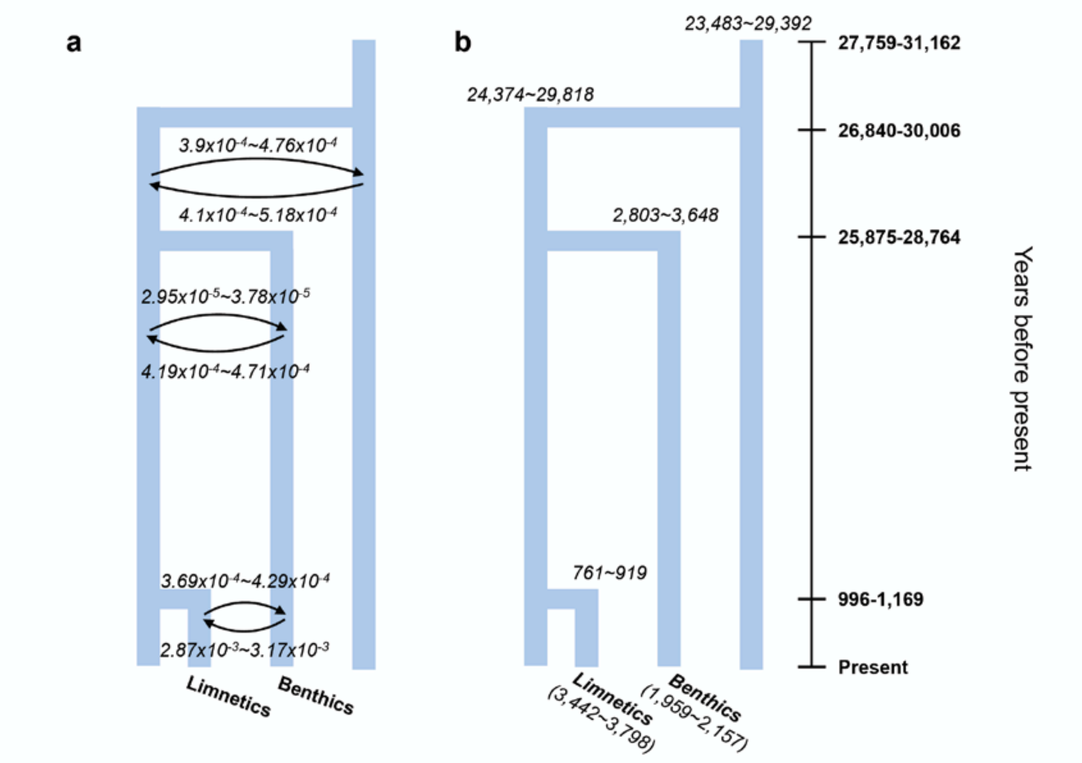

Figure 4.7 | Demographic model of Paxton Lake benthics and limnetics inferred by δaδi. All the ranges correspond to 95% confidence intervals from 100 conventional bootstraps. **a,** migration rates of gene flow events between different ancestral/recent populations. **b,** divergence time and population size of different ancestral/recent populations. The divergence time is denoted to the left of the plot.

4.4 Discussion

4.4.1 Genetic relationship and ancestry of benthics and limnetics

Coyne & Orr (2004) proposed three criteria for identifying sympatric speciation: 1) overlapping habitat, 2) speciation must be complete, 3) species arise from sympatric speciation should be sister groups or monophyletic cluster(Coyne & Orr 2004). Benthics and limnetics reside in same lakes and have overlapping habitat. In addition, previous studies identified reproductive isolation between benthics and limnetics(McPhail 1993). Thus, the most important evidence for sympatric speciation of benthics and limnetics is whether species-pairs from the same lake form a monophyletic group(Coyne & Orr 2004). Previous phylogenetic analysis of benthics and limnetics using microsatellite markers revealed species pairs from the same lake formed polyphyletic groups, which is consistent with the prediction of allopatric speciation. However, the phylogenetic trees generated by this study are ambiguous due to the limited number of markers (Taylor & McPhail 1999). The phylogenetic tree of benthics and limnetics as well as marine and freshwater inferred in this thesis study using whole-genome SNPs demonstrated benthics and limnetics from all four lakes formed distinct clades respectively. This suggests the species pair from the same lake did not derive from a common ancestral population.

PCA of benthics/limnetics and 210 marine/freshwater sticklebacks using genome-wide SNPs separates benthics and limnetics by species on PC1 and by lakes on PC2, implying benthics or limnetics from different lakes have closer relationship than species pair from the same lake. PCA using genome-wide SNPs places freshwater sticklebacks closed to benthics and marine sticklebacks closed to limnetics. This suggests limnetics have a closer genetic relationship with marine sticklebacks than benthics, while benthics are genetically close to freshwater sticklebacks, which is consistent with the prediction of double-invasion hypothesis and the result of previous study (Jones et al 2012a). Conversely, benthics and limnetics from the same lake cluster in the PCA using neutral SNPs. It indicates a close genetic relationship of benthics and limnetics from the same lake at neutral regions, which may arise from gene flow between the species pair.

Inferring the genetic relationship of benthics and limnetics in the context of a large set of marine and freshwater sticklebacks allows me to investigate the ancestry of these two species. Benthics from all four lakes cluster with freshwater ecotypes from the nearby Little Campbell River (LITC_UP) in the conventional maximum likelihood (ML) phylogenetic trees generated based on sequence divergence and the TreeMix ML tree constructed based on allele frequency. In addition, PCA places LITC_UP close to benthics from all four lakes. This suggests benthics and LITC_UP have a close genetic relationship. The analysis of the ancestry of benthics using outgroup f_3 statistic indicates benthics share most ancestry with LITC_UP. In contrast, limnetics from three lakes (except Enos Lake) formed a monophyletic clade in the conventional ML tree and do not cluster with other marine and freshwater populations. Furthermore, the analysis using outgroup f_3 statistics cannot identify a clear population that have shared ancestry with limnetics from three lakes (except Enos Lake). This can be resulted from: 1) the population that share ancestry with limnetics is not sampled and analyzed in this study, 2) the unique evolutionary history of limnetics after they diverged from the ancestral population, 3) gene flow from benthics to limnetics.

Enos Lake benthics and limnetics formed a monophyletic clade in conventional ML tree, which is consistent with the prediction of sympatric speciation. However, PCA using genome-wide SNPs places Enos Lake limnetics between the benthics and limnetics clusters, which might suggests monophyletic clustering of Enos Lake benthics and limnetics is due to the increased gene flow between them. In addition, although the results of outgroup f_3 test for Enos Lake limnetics are more similar to those of its species pair (Enos Lake benthics) compared to other species pairs, Enos Lake benthics and limnetics have clearly different test result. This indicates they do not have common ancestor and suggests the close phylogenetic relationship between Enos Lake species pair is because of increased gene flow rather than sympatric speciation.

4.4.2 Improved demographic model of Paxton Lake benthics and limnetics

δaδi infer the common ancestral population of Paxton Lake benthics and limnetics diverged from an ancestral population between 28,640 to 30,006 years ago (95% C.I.), and SMC++ infers both Paxton Lake benthics and limnetics have experienced a population bottleneck between 20,000 to 30,000 years ago. This suggests there might be a split of ancestral marine population starting at 30,000 years ago, and Paxton Lake benthics and limnetics were derived from one of the resulting populations. δaδi infers the ancestral population of benthics diverged from the common ancestral population between 25,875 and 28,764 years ago (95% C.I.), which is very closed to the time when the common ancestral population diverged from its ancestors. The demographic analysis using SMC++ infers a recent population size decline of Paxton Lake benthics and limnetics at about 7,000 and 5,000 years ago, which should be correspond to the times of colonization of the Paxton Lakes by the ancestors of benthics and limnetics separately. This is a direct genetic evidence of the double-invasion hypothesis, which proposed the ancestors of benthics and limnetics colonized the lake separately in about 1,500 years. δaδi infers Paxton Lake limnetics diverged from the common ancestors between 996 to 1169 years ago, and SMC++ infers Paxton Lake limnetics reach the bottom of the recent population size decline (starts at 5,000 years ago) at about 1,500 years ago. This suggests after the colonization of Paxton Lake, the gene flow between the ancestors of benthics and limnetics is high. The gene flow between species started to decrease and the reproductive isolation gradually accumulated due to divergent natural selection. The reproductive isolation between Paxton Lake benthics and limnetics formed at about 1,000 years ago. Paxton Lake benthics reached the bottom of the recent population size decline about 500 years later than limnetics, which may due to the stronger natural selection acted on Paxton Lake benthics. Taken together, I hypothesize a demographic model of Paxton Lake benthics and limnetics as illustrated in **Fig. 4.8**.

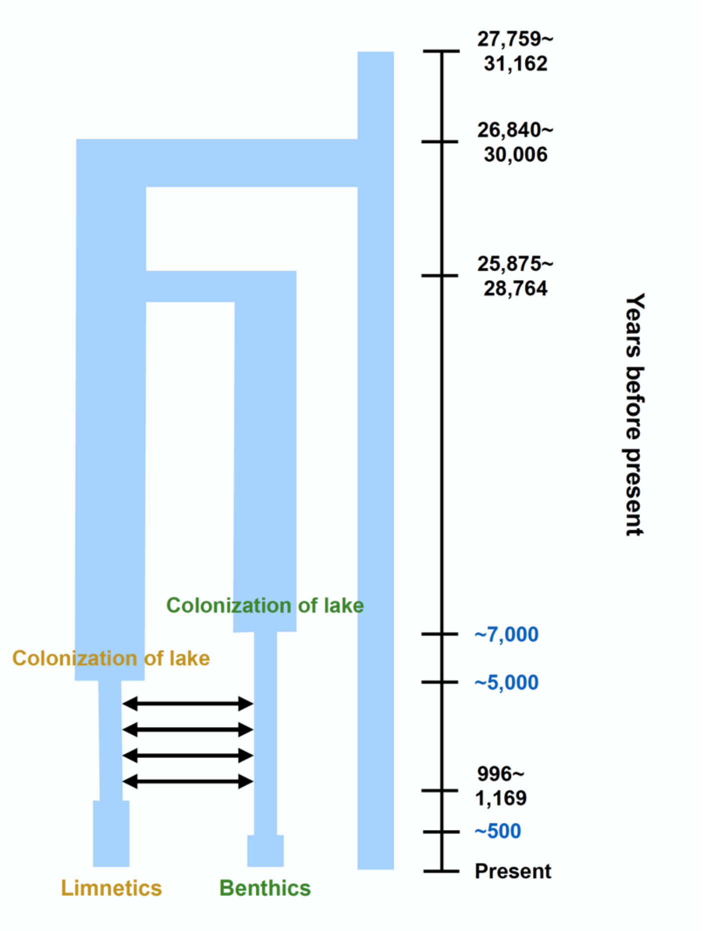

Figure 4.8 | Improved demographic model of Paxton Lake benthics and limnetics.

4.5 Materials and Methods

4.5.1 Principal Component Analysis (PCA)

To elucidate the evolutionary history of benthics and limnetics, the genetic relationship of benthics, limnetics and global marine and freshwater sticklebacks was assessed using PCA. PCA was performed using smartpca program v13050 with genome-wide SNPs (Patterson et al 2006), SNPs in the neutral regions separately. As the genetic divergence between Pacific and Atlantic populations is large, PCA analyses were first performed for benthic and limnetic individuals, and marine and freshwater stickleback individuals were projected onto the PC space of benthics and limnetics. For PCA using whole-genome variants, 6,134,540 SNPs were used in the analysis after filtering by smartpca program. To eliminate the effect of selection, the SNPs with high degree of linkage disequilibrium (LD) were removed using the LD correction function of smartpca program with option "`nsnpldregress 2`".

SNPs in the genomic regions having *P*-value larger than 0.5 in the permutation analysis of CSS scores in cross-lake benthics and limnetics were identified as neutral SNPs. In total, 15,100,514 SNPs from 8,681 neutral genomic regions were inputted into smartpca program. After filtering, 5,761,616 SNPs were used for PCA analysis.

4.5.2 Phylogenetic and genetic distance relationship analysis

The phylogenetic tree of benthics, limnetics and global marine and freshwater stickleback individuals was constructed using whole genome genetic variants. To eliminate the effect of rare variations, the SNPs dataset was filtered for SNPs with minor allele frequency less than 0.01 using VCFtools v0.1.14. The SNPs were concatenated into consensus sequence for each individual using custom Python script. The phylogenetic tree was estimated using 9,012,726 SNPs for 258 stickleback individuals. Due to the computational limitation, I first estimated the maximum-likelihood (ML) phylogenetic tree using RAxML (Stamatakis 2014) v8.1.20 under GTRGAMMA nucleotide substitution model. Approximately-maximum-likelihood tree was constructed with FastTree (Price et al 2010) v 2.1.10 using the ML tree estimated by RAxML as starting tree. The tree was constructed using GTR+CAT approximation model with 20 rate categories. The tree was annotated in dendroscope program (Huson et al 2007) v3.5.9.

Admixture among stickleback populations was modeled using TreeMix v1.12 (Pickrell & Pritchard 2012). TreeMix analysis was performed for benthics and limnetics as well as marine and freshwater stickleback populations. To eliminate SNP calling errors due to low coverage or mapping errors, SNP sites with mean depth of coverage less than 3X or more than 100X were removed using VCFtools v0.1.14. In total, 13,778,805 SNPs were inputted into TreeMix for the analysis.

4.5.3 Ancestry of benthics and limnetics

To evaluate the pattern of admixture and shared ancestry between benthics/limnetics and marine/freshwater stickleback populations, I calculated outgroup f_3 statistic using qp3Pop program v300 implemented in EIGENSOFT package. Outgroup f_3 was calculated for benthics/limnetics and marine/freshwater stickleback populations with more than 4 individuals in benthics, limnetics and global marine and freshwater stickleback variants dataset using marine population from River Tyne as outgroup. To avoid SNP calling errors due to low coverage or alignment errors, I filtered out SNP sites with mean depth of coverage less than 3 or more than 100 as well as sites with missing genotype calls more than 80% using VCFtools v0.1.14. The results were plotted using custom R script.

4.5.4 Demographic inference of Paxton Lake benthics and limnetics

4.5.4.1 SMC++

Historical effective population sizes of benthics and limnetics was inferring using smc++ v1.11.0 (Terhorst et al 2016). To eliminate the effect of selection and retain the complete pattern of LD, I used SNPs from the putative “neutral” chromosome (chrXV) which has no QTL mapped in benthics and limnetics for several phenotypic traits (Arnegard et al 2014, Conte et al 2015) for the analysis. Ancestral allele of each SNP site was determined as the major allele of marine ecotypes from Little Campbell River and River Tyne. To avoid the SNP calling errors due to the alignment error, SNPs located in the previously identified centromeric repeats (Cech & Peichel 2015) and repetitive regions (Jones et al 2012b) were filtered from the dataset. Effective population size was inferred with mutation rate of $6\text{x}10^{-8}$ estimated by mlRho. Historical effective population size of Paxton Lake benthics or limnetics was estimated using genotypes of 23 individuals The history of population size was plotted with average generation time of 1 year using custom R script.

4.5.4.2 δaδi

Twenty-three Paxton Lake benthics and 23 Paxton Lake limnetics were used for the demographic inference using δaδi. To remove the effect of selection, 2,667,791 SNPs from genomic regions that are not diverged between Paxton Lake benthics and limnetics (CSS, *P*-value > 0.5) were used in the analysis. Four demographic models of Paxton Lake benthics and limnetics were evaluated using δaδi program and the model with highest Poisson likelihood were used to estimate demographic parameters. To obtain confidence intervals for the estimate of each parameter, 100 bootstrap datasets were generated using custom Python script. The parameters were inferred for each bootstrap dataset and used to construct confidence intervals.

5 Gene expression divergence of benthics and limnetics

5.1 Background and Aims

Besides the evolution of variations in gene sequences, the evolution of gene expression due to regulatory sequence divergence plays important role to the phenotypic diversity in nature (King & Wilson 1975, Stern & Orgogozo 2008, Stern & Orgogozo 2009). The interaction of *cis*- and *trans*-regulatory elements regulate the expression of target gene (Stern & Orgogozo 2009). *Cis*-regulatory elements are physically linked on the same DNA molecule to the genes whose expression they regulate, and *trans*-regulatory factors can control expression of genes that are distant from which they were transcribed (Mack & Nachman 2017). It has been argued that *cis*-regulation is particularly important for phenotypic evolution because it provides a mechanism for spatial and temporal fine tuning of gene expression via mutations in non-coding regulatory modules that avoids causing amino acid changes and their potentially deleterious pleiotropic effects (Prud'homme et al 2007). Further, natural selection is thought to be more efficient at filtering *cis*-regulaory than *trans*-regulatory elements because they are directly linked to the genes whose expression they regulate and are more rapidly purged from the population if they have deleterious effects on gene expression (Wittkopp & Kalay 2012, Wray 2007a).

Cis-regulatory divergence of gene expression can be inferred in interspecific crosses from the observation of allele-specific expression (Pastinen 2010). A diploid individual carries alleles from each of its parents which can often be distinguished from each other by the presence of polymorphisms. A null expectation is that within a given individual both maternal and paternal versions of the gene are transcribed at equal levels. However expression is often biased towards either maternal or paternal allele – a phenomena called *allele specific expression* (ASE) (Pastinen 2010). ASE analysis quantifies the expression levels of maternal and paternal transcripts (Yan et al 2002). Since the trans-acting environment within the nucleus is the same for both maternal and paternal chromosomes, any allele-specific expression can only be attributed to differences in the *cis*-regulatory landscape (Pastinen 2010).

Dissecting the role of *cis*-regulation in gene expression has greatly improved our understanding of gene expression evolution in several species (Goncalves et al 2012, He et al 2012, Wang et al 2017). The study of expression divergence between *Drosophila melanogaster* and *Drosophila simulans* showed 28 out of 29 test genes showed *cis*-regulatory divergence (Wittkopp et al 2004). In addition, the study of differential allelic gene expression between and within *Drosophila* species (*D. melanogaster*, *D. simulans*) revealed *cis*-regulatory changes accounted for greater proportion of expression difference between than within species, suggesting natural selection plays a role in divergent gene expression (Wittkopp et al 2008). Genomic analysis gene expression divergence between two yeast species demonstrated expression is largely attribute to *cis*-regulatory divergence in stable conditions, while *trans*-regulatory divergence contributes to the rapid response to environmental changes (Tirosh et al 2009).

Recent studies showed phenotypic divergence between marine and freshwater sticklebacks were due to divergent expression of adaptive genes mediated by changes in nearby *cis*-regulatory elements (Chan et al 2010, Cleves et al 2014, Miller et al 2007, O'Brown et al 2015). In addition, genome-wide gene expression divergence between marine and freshwater sticklebacks was predominantly attributed to *cis*-regulatory changes (Verta & Jones). This suggests *cis*-regulation changes play an important role in the adaptation of sticklebacks. However, the regulation of gene expression in the sympatric benthics and limnetics and the role of *cis*-regulatory changes to their speciation are largely unknown. Since the phenotypic divergence of benthics and limnetics involves multiple different phenotypic and behavioral traits with independent genetic basis (Arnegard et al 2014, Conte et al 2015), the *cis*-regulatory hypothesis therefore predicts that adaptive divergence is mediated by multiple *cis*-regulatory changes with a dispersed genomic distribution. Using an allele-specific expression assay I tried to quantify the role of *cis*-regulation of gene expression in the divergence of benthics and limnetics.

In this chapter, I identified genome-wide pattern of *cis*-regulatory divergence of Paxton Lake benthics and limnetics using F1 hybrids. The objectives of this chapter are:

- to identify genes that show *cis*-regulatory divergence of expression in Paxton Lake benthics and limnetics
- to evaluate the biological functions and determine the selective pattern of genes showing *cis*-regulatory divergence of expression.

5.2 Allele-specific expression analysis of Paxton Lake benthics and limnetics

5.2.1 Study samples and sequencing

Allele-specific expression (ASE) analysis was performed using F1 hybrids of wild-caught Paxton Lake benthics and limnetics. Two F1 families each of reciprocal crosses of Paxton Lake benthics and limnetics (benthics x limnetics, limnetics x benthics) were generated in the wild and shipped to the stickleback fish facility at the Max Planck Institute for Developmental Biology in Tübingen. The F1 individuals were reared under common garden standard husbandry condition until they were 30 days post fertilization. Fishes were then euthanized and RNA sequencing (RNA-Seq) libraries were prepared from whole bodies using Illumina RNA-Seq library construction kit. RNA-Seq was performed for all the F1 individuals using standard Illumina 2x150bp chemistry.

As ASE analysis dissects patterns of allele specific expression using allelic polymorphisms within the transcribed gene, whole genome DNA sequencing (Illumina 2x150bp) was performed for the parental fish of all four F1 crosses and sites where parents were homozygous for alternate alleles were identified. High-confidence fully-informative SNPs (parents are homozygous for alternate alleles at this position) account for ~20% of total SNPs identified in parental fishes of each F1 cross (**Table 5.1**). The distance between informative SNPs is high (~500bp) (**Table 5.1**), which facilitates the ASE analysis of Paxton Lake benthics and limnetic.

Table 5.1 Information of informative SNPs in parents of F1 families

Parent (Female)	Parent (male)	SNPs	Informative SNPs	Proportion	Distance between informative SNPs (bp)
Benthic_7	Limnetic_7	4,069,401	774,706	19%	598
Benthic_8	Limnetic_8	4,080,982	923,556	22.6%	501
Limnetic_10	Benthic_10	4,068,750	899,109	22%	515
Limentic_11	Benthic_11	4,073,936	972,049	23.9%	476

5.2.2 Transcriptome assembly

The stickleback genome has a high quality gene-set annotation performed by Ensembl that is based on gene predictions from the reference assembly (freshwater ecotype) combined with information on expressed genes derived from the sequencing of marine and freshwater expressed sequence tags (ESTs) libraries from multiple tissues and individuals (>350,000 sequenced clones) (Jones et al 2012b). The latest version of this gene-build (v90) has 29,044 transcript predictions arising from 22,442 genes (Zerbino et al 2018). The number of predicted coding genes is similar to other well-annotated gene builds (sticklebacks: 20,787; fugu: 18,523; human: 20,805; mouse: 23,148; *C.elegans*: 20,532), however the number of transcripts is considerably lower (sticklebacks: 29,245; fugu: 48,706; humans: 196,501; mouse: 94,647; *C. elegans:* 57,844) (Zerbino et al 2018). Rather than a biological absence of transcript splice variants in sticklebacks, this relatively low transcript count is more likely due to lack of data. Improving transcript annotation can therefore aid studies of gene expression in sticklebacks, as gene expression has the potential to play a significant role in evolutionary adaptation to different environments.

Because benthic and limnetic ecotypes are likely to have diverged in their transcriptome relative to each other, and the gene annotations were performed for the freshwater reference genome, I performed a reference-guided transcriptome assembly based on RNA-Seq data from all F1 individuals. First, individual transcriptome assemblies were made using STARR aligner and Cufflinks with the reference genome transcripts used as a guide. Then transcriptome assemblies of individuals from each F1 cross (BL_7, BL_8, LB_10, LB_11) were merged using Cuffmerge (**Table 5.2**).

Furthermore, transcriptome assemblies of all F1 individuals were merged using Cuffmerge (BenLim merged).

Table 5.2 Summary of gene prediction for Paxton Lake benthic-limnetic F1 crosses

	Ensembl (v90)	BL_7	BL_8	LB_10	LB_11	BenLim merged
Gene	22,442	24,274	24,424	24,365	24,345	24,482
Transcript	29,044	68,308	68,666	68,260	68,633	107,351

Note: BL_7 and BL_8 are two F1 families with direction benthics x limnetics, and LB_10 and LB_11 are two F1 families with direction limnetics x benthics

5.2.3 Allele Specific Expression (ASE) analysis

ASE was quantified in F1 individuals from each of 4 independent benthic x limnetic crosses (2 x each reciprocal direction) by placing RNA-Seq reads against the assembled transcriptome, identifying reads that fall within transcripts and span fully informative SNPs, and comparing expression levels of the alternate alleles. Four individuals from each F1 cross were used for ASE analysis to eliminate the effect of genetic variations between cross parents. As most of the genes have multiple predicted transcripts with different length, the presence/absence and the number of informative SNPs located can vary among different transcripts of a gene. Therefore, I used the longest transcript of each gene in the ASE analysis. More than half of genes (~60%) contain at least one informative SNP, and therefore used in ASE analysis (**Table 5.3**).

Table 5.3 Summary of genes used for ASE analysis

	BL_7_1	BL_7_2	BL_7_3	BL_7_4
Total	7,267	7,384	7,020	6,874
Proportion	58.4%	59.4%	56.4%	55.3%
	BL_8_1	**BL_8_2**	**BL_8_3**	**BL_8_4**
Total	8,234	8,172	8,015	8,283
Proportion	61.8%	61.4%	60.2%	62.2%
	LB_10_1	**LB_10_3**	**LB_10_4**	**LB_10_5**
Total	8,917	8,776	8,809	8,610
Proportion	64.9%	63.9%	64.1%	62.7%
	LB_11_1	**LB_11_2**	**LB_11_3**	**LB_11_4**
Total	8,747	8,820	8,639	8,759
Proportion	62.4%	63%	61.7%	62.5%

ASE was tested for each informative SNP site in each F1 individual using binomial exact test with FDR level of 10%. About 2,000 genes contained at least one significant ASE SNP sites, suggesting expression divergence of them between Paxton Lake benthics and limnetics may be *cis*-acting (**Table 5.4**). These genes account for ~10% of total and ~20% of the analyzed genes in the genome (**Table 5.4**), which is similar to the proportion of *cis*-regulatory diverging genes of marine and freshwater populations from Little Campbell River (Verta & Jones). This suggests *cis*-acting divergence is also prevalent in Paxton Lake benthics and limnetics, and might play an important role in their adaptation and speciation.

Table 5.4 Summary of putative *cis*-regulatory diverging genes in F1 individuals of Paxton Lake benthics and limnetics

	BL_7_1	**BL_7_2**	**BL_7_3**	**BL_7_4**
Number of Genes	2,262	1,550	1,618	1,176
Proportion (total genes)	9.3%	6.4%	6.6%	4.8%
Proportion (analyzed genes)	31.1%	20.9%	23%	17%
	BL_8_1	**BL_8_2**	**BL_8_3**	**BL_8_4**
Number of Genes	2,390	2,503	2,214	1,912
Proportion (total genes)	9.7%	10.2%	9%	7.8%
Proportion (analyzed genes)	29%	30.6%	27.6%	23.1%
	LB_10_1	**LB_10_3**	**LB_10_4**	**LB_10_5**
Number of Genes	2,143	2,333	2,132	2,561
Proportion (total genes)	8.7%	9.6%	8.7%	10.5%
Proportion (analyzed genes)	24%	26.50%	24.20%	29.7%
	LB_11_1	**LB_11_2**	**LB_11_3**	**LB_11_4**
Number of Genes	2,156	2,185	2,387	2,676
Proportion (total genes)	8.80%	8.90%	9.80%	10.90%
Proportion (analyzed genes)	24.60%	24.70%	27.60%	30.50%

As there are genetic polymorphisms within each parental fish, full-sib offspring may not necessarily inherit the same alleles. As a consequence, the genes that have allele-specific expression are not highly correlated among individuals from the same F1 cross (**Appendix Table 14**). To identify genes that showed *cis*-regulatory divergence in parallel between Paxton Lake benthics and limnetics, I filtered the genes having at least one significant ASE SNP with two criteria: first, the gene has to have one significant ASE SNP and at least one SNP site have expression difference with same direction as the

ASE SNP; second, the gene has to show ASE in all four individuals of F1 cross. The second criterion is very conservative and may increase the false negative rate, but it can identify genes showing *cis*-regulatory divergence in the various genetic backgrounds. These genes are likely to contribute to the phenotypic divergence between Paxton Lake benthics and limnetics during their adaptation. In total, 762 and 888 *cis*-regulatory diverging genes were identified in crosses with mating direction benthics x limnetics and limnetics x benthics respectively. These genes were considered as having consistent *cis*-regulatory divergence between Paxton Lake benthics and limnetics, and used in the subsequent analyses.

5.3 Functions of gene with *cis*-regulatory divergence between Paxton Lake benthics and limnetics

I performed GO enrichment analysis to determine the function of *cis*-diverging genes of Paxton Lake benthics and limnetics. As zebrafish has the best GO annotation of fish species and better syntenic relationship with sticklebacks, GO enrichment analyses were performed using zebrafish orthologs of stickleback *cis*-regulatory diverging genes. *Cis*-regulatory diverging genes identified in reciprocal crosses showed significant enrichment in biological processes of muscle cell development, carbohydrate catabolic process, inner ear/otolith development, heart development, ion transport, organ morphogenesis, and fin regeneration (**Table 5.5**). It is note worthy that genes involved in muscle development, cardiovascular development, anatomical/organ morphogenesis and ion transport are also significantly enriched in GO enrichment analyses of genes in “composite adaptive regions” of benthics and limnetics (see **Section 3.3.2**). This suggests these biological processes are critical for the adaptation of benthics and limnetics, and expression divergence of genes involved in these processes is *cis*-acting. Benthics and limnetics have large phenotypic divergence due to the adaptation to different environments in less than 13,000 years (McPhail 1994). The divergence in phenotypic traits of benthics and limnetics requires divergence at several genomic regions derived from directional selection. GO enrichment analysis showed genes in adaptive regions are significantly enriched in several important biological processes of fish development and survival (**see Section 3.3.2**). Genes involved in these processes would have functional constraint, and thus, genetic changes are more likely to appear and fix in their *cis*-regulatory elements during rapid adaptation.

Otolith development are significantly enriched in GO enrichment analyses of *cis*-regulatory diverging genes and genes containing benthic-specific variants (see **Section 3.5**). This suggests expression divergence of genes regulating otolith development between Paxton Lake benthics and limnetics have a *cis*-acting basis, and infers a phenotypic divergence of otolith in the sympatric species pair. Otolith is a calcium carbonate structure in the inner ear of all vertebrate species (Sheykholeslami & Kaga 2002). Saccule, utricule, and lagena are three otolith organs that help fishes to detect sounds and linear acceleration under water (Popper et al 2005, Webb et al 2006). Adaptive variations of inner ear, especially of saccular otolith, have been documented in different teleost fish species, which facilitate adaptation of fish to different environmental niches (Cruz & Lombarte 2004, Lombarte & Cruz 2007). The growth of saccular otolith in fish has a genetic basis, and is influenced by the decline of temperature due to the increment of water depth (Lombarte & Lleonart 1993). Therefore, divergence of otolith development between benthics and limnetics might due to the adaptation to the depth and temperature variations of their habitats.

In total, 61 *cis*-regulatory diverging genes are located in the “composite adaptive regions” of benthics and limnetics (**Table 5.6**), and these genes are predominantly involved in important biological processes including anatomical development, cardiovascular system development, ion transport, and eye development. This suggests these biological processes are critical for the adaptation of benthics and limnetics, and expression divergence of these genes has a *cis*-acting basis.

Table 5.5 Enrichment of Gene Ontology categories of *cis*-regulatory diverging gene

GO category	Annotated	Observed	Expected	*P*-value	Genes Included
muscle cell development	76	11	4.04	0.0021	MTM1, NEB, CAV3, OGG1, MYOD1, TCAP, GSK3AB, TNNT2A, MEF2AA, ENSGACG00000004227, ENSGACG00000015181
carbohydrate catabolic process	41	7	2.18	0.0054	GAPDHS, ENO1A, TPI1B, GAPDH, ENSGACG00000009411, ENSGACG00000020677
inner ear development	85	9	4.52	0.0361	EYA1, SOX2, DFNA5B, CEP290, HSP90B1, TGFB1A, ARHGEF11, ATP1B2B, ENSGACG00000018016
otolith development	23	5	1.22	0.0063	EYA1, HSP90B1, TGFB1A, ARHGEF11, ATP1B2B
heart development	214	20	11.39	0.0099	OGG1, GSK3B, GSK3AB, CSAD, TNNT2A, KRAS, DPF3, MEF2AA, ATP2A2A, RP2, MMD, NPNT, BNIP3LB, PDCD4B, RASSF8B, RUVBL2, YWHAG1, ATP1B2B, ENSGACG00000004227, ENSGACG00000016114
Golgi vesicle transport	27	5	1.44	0.0127	GOSR1, COG5, SEC24D, ENSGACG00000004658, ENSGACG00000013339
dephosphorylation	111	12	5.91	0.0146	MTM1, CA16B, EYA1, PTPRC, RNGTT, PPP1R2, PPM1NA, PPM1G, SBF1, PTP4A1, SYNJ1, PPM1E
Animal organ morphogenesis	322	26	17.14	0.0216	EYA1, GSK3B, EIF3EA, RBCK1, SOX2, DFNA5B, PAFAH1B1A, CEP290, GSK3AB, TNNT2A, IFT80, ITGA5, KRAS, DPF3, ATP2A2A, RP2, MMD, PMM2, NPNT, BNIP3LB, SEC24D, PLCB3, ATP1B2B, SYNJ1, ENSGACG00000004227, ENSGACG00000005658
ion transport	437	33	23.26	0.0255	SLC4A8, SLC39A10, ARMC1, CA1, SLC25A22, CLIC4, CACNA1SB, ITPR2, SLC7A4, CLIC5B, KCTD13, RYR1B, ATP2A2A, ABCB11B, ATP1B1B, ATP11C, SLC30A9, GRIA1A, SLC7A2, CLCN3, SLC8A4B, CHRNB1, ATP2A3, GRIA3A, ATP1B2B, SLC9A6B, SLC13A2, GABRA1, ENSGACG00000000423, ENSGACG00000001024, ENSGACG00000001755, ENSGACG00000005658, ENSGACG00000007545
chloride transport	23	4	1.22	0.0313	CLIC4, CLIC5B, CLCN3, GABRA1
sodium ion transport	43	6	2.29	0.0251	CA2, ATP1B1B, ATP1B2B, SLC9A6B, SLC13A2, ENSGACG00000007545
myofibril assembly	32	5	1.7	0.0256	CAV3, TNNT2A, MEF2AA, ENSGACG00000004227, ENSGACG00000015181
fin regeneration	33	5	1.76	0.0289	KRT5, SOX2, HAPLN1A, ANXA1A, CTSBA

Table 5.6 Functions of adaptive genes with *cis*-regulatory divergence

No.	Ensembl Gene ID	Gene Name	Zebrafish Gene Ontology Annotation	Zebrafish Knockdown Phenotype	Reference
1	ENSGACG00000000644	*SUMF1*			
2	ENSGACG00000000663				
3	ENSGACG00000000758				
4	ENSGACG00000000858				
5	ENSGACG00000000872	*USP4*	negative regulation of toll-like receptor signaling pathway, protein ubiquitination	eye decreased size, head decreased size, notochord development disrupted	(Tse et al 2009)
6	ENSGACG00000000913	*CAV3*	caveola assembly, muscle cell fate commitment, notochord cell development, sarcomere organization	notochord inner cell collapsed	(Garcia et al 2017)
7	ENSGACG00000001211	*KAZNB*			
8	ENSGACG00000001254	*TWF2A*	negative regulation of actin filament polymerization		
9	ENSGACG00000001501	*FRS3*			
10	ENSGACG00000005546	*NT5C*			
11	ENSGACG00000007288	*IVNS1ABPA*			
12	ENSGACG00000007546				
13	ENSGACG00000007569	*HSC70 / HSPA8*	fin regeneration, positive regulation of receprot-mediated endocytosis, regulation of fibroblast growth factor receptor signaling pathway	ceratohyal cartilage deformed, hindbrain decreased size, Meckel's cartilage deformed, eye decreased size	(Amsterdam et al 2004, Robu et al 2007)
14	ENSGACG00000007733	*CSNK1G2A*	endocytosis, protein phosphorylation, Wnt signaling pathway		
15	ENSGACG00000008448	*TLCD2*			
16	ENSGACG00000008536	*MLKL*			
17	ENSGACG00000008820				
18	ENSGACG00000008901	*PHKG1A*	angiogenesis, glycogen biosynthetic process, phosphorylation	angiogenesis decreased process quality	(Camus et al 2012)

19	ENSGACG00000009210	*TES*	regulation of cell proliferation		
20	ENSGACG00000009214	*CPA1*	proteolysis		
21	**ENSGACG00000009278**	***OPA1***	**apoptotic process, chordate embryonic development, mitochondrial fusion**	**cardiac muscle cell increased size, blood circulation disrupted, eye decreased size**	**(Li et al 2014a, Rahn et al 2013)**
22	**ENSGACG00000009295**	***ATP13A3***	**cation transport**		
23	**ENSGACG00000009373**	***Kitlg***		**Regulate gill and ventrum pigmentation in sticklebacks**	**(Miller et al 2007)**
24	ENSGACG00000009446	*SLC1A5*			
25	ENSGACG00000009469	*EGLN2*			
26	**ENSGACG00000009747**	***TNNT2A***	**artery development, blood circulation, heart contration, muscle contraction**	**blood circulation decreased rate. heart contraction arrested**	**(Chen et al 2017)**
27	ENSGACG00000009752	*PKP1 (1 of many)*			
28	ENSGACG00000010554	*FAM120C*			
29	ENSGACG00000010685	*LIMA1A*	actin filament bundle assembly, negative regulation of actin filament depolymerization	forebrain/hindbrain, midbrain lima1a expression decreased amount	(Jungke et al 2016)
30	ENSGACG00000010714				
31	**ENSGACG00000011015**	***SOCS3A***	**regeneration, cytokine-mediated signaling pathway, posterior lateral line neuromast hair cell development, retina morphogenesis**	**Posterior lateral line neuromast decreased amount**	**(Liang et al 2012)**
32	ENSGACG00000018533				
33	ENSGACG00000018752	*HDAC3*	angiogenesis, liver development, histone deacetylation, covalent chromatin modification	liver decreased size, posterior lateral line neuromast abnormal	(Farooq et al 2008, He et al 2016)
34	ENSGACG00000019116	*SLC16A7*			
35	ENSGACG00000019333	*ALDH1L2*	10-formyltetrahydrofolate catabolic process, oxidation-reduction process		
36	ENSGACG00000019336	*SLC41A2B*	cation transport, transmembrane transport		

37	ENSGACG00000019457	*PHLDA1*			
38	ENSGACG00000019459	*NAP1L1*	nucleosome assembly, response to yeast		
39	ENSGACG00000019461	*OSBPL8*			
40	ENSGACG00000019943	*CHRNB1*	cation transport, ion transport	eye decreased size, head decreased size, liver hypoplastic	(Amsterdam et al 2004)
41	ENSGACG00000019950				
42	ENSGACG00000020023				
43	ENSGACG00000020024				
44	ENSGACG00000020072	*EIF4A1A*	protein desumoylation		
45	**ENSGACG00000020152**	***SERPINH1A***	**collagen fibril organization**		
46	**ENSGACG00000020236**				
47	**ENSGACG00000020239**	***MTMR4***	**dephosphorylation, transforming growth factor beta receptor signaling pathway**		
48	ENSGACG00000020257				
49	**ENSGACG00000020259**	***CENPV***	**metabolic process**		
50	**ENSGACG00000020260**	***NCOR1-like***	**Anterior/posterior pattern specification, hindbrain development, neutrophil differentiation**	**Neutrophil decreased amount, anterior/posterior pattern specification disrupted, hindbrain dcreased length**	**(Li et al 2014b, Xu et al 2009)**
51	ENSGACG00000020265				
52	**ENSGACG00000020333**	***MSNA***	**blood vessel lumenization, endoderm development**	**blood vessel lumenization process quality abnormal, intersegmental vessel unlimenized**	**(Wang et al 2010)**
53	ENSGACG00000020353	*PPME1*	protein demethylation		
54	ENSGACG00000020354	*UCP3*	adaptive thermogenesis, mitochondrial transmembrane transport, response to cold		
55	ENSGACG00000020359	*PAFAH1B2*	brain development		
56	ENSGACG00000020360	*MPZL2B*	cell-cell adhesion		
57	ENSGACG00000020394	*MINK1*	actin cytoskeleton reorganization, neuron projection morphogenesis, protein phosphorylation		

58	ENSGACG00000020395	*GNB2*	signal transduction		
59	ENSGACG00000020398	*GUCY2D*	cGMP biosynthesis process, cyclic nucleotide biosynthetic process, protein phosphorylation	visual behavior quality abnormal, visual perception quality abnormal, optomotor response arrested	(Muto et al 2005)
60	ENSGACG00000020400	*SLC25A15B*	Mitochondrial ornithine transport		
61	ENSGACG00000020404	*SPTBN2*	actin filament capping		

Note: *Cis*-diverging genes located in “strongly adaptive regions” are highlighted in red

Interestingly, within the 61 *cis*-regulatory diverging genes that are located in “composite adaptive regions”, 11 are located in “strongly adaptive regions” of benthics and limnetics (**Table 5.6**), indicating they were subjected to divergent selections in both benthics and limnetics during their adaptation. It is noteworthy that *Kitlg* gene, which regulates gill and ventrum pigmentation in Paxton Lake benthics and limnetics, showed significant allele specific expression in the analysis. This is consistent with the result of previous study (Miller et al 2007). Furthermore, one adaptive gene (*SOCS3*) that was studied in previous chapter (see **Section 3.3.1**) has *cis*-regulatory divergence between Paxton Lake benthics and limnetics. *SOCS3* (chrXI: 9,066,121-9,067557) forms a negative feedback loop with *STAT3*, and regulates tissue regeneration and neuromast development in zebrafish (Liang et al 2012). The downstream intergenic region of *SOCS3* is highly divergent between cross-lake benthics and limnetics. Additionally, the intergenic region has been subject to strong divergent selection in Paxton Lake benthics and limnetics. Allele-specific expression of *SOCS3* further suggests the *cis*-regulatory divergence of *SOCS3* may play an important role in the adaptation of benthics and limnetics. Sequence comparison showed there was a deletion (chrXI: 9,055,533-9,058,908) ~7kb downstream of *SOCS3* in Paxton Lake benthics but not in limnetics, which is experimentally confirmed (**Fig. 5.1**). Analysis of the intergenic region in benthics and limnetics from other lakes showed the deletion was fixed in benthics. Interestingly, the deletion overlaps with a long interspersed nuclear element-1 (LINE-1). It indicates the deletion removed the LINE-1 retrotransposon from the intergenic region of *SOCS3* in Paxton Lake benthics. It has been showed that LINE removal from the regulatory sequence of a gene can affect its expression, which further causes phenotypic divergence in vertebrates (Bohne et al 2008, Elbarbary et al 2016). Thus, *cis*-regulatory divergence of *SOCS3* might attribute to the deletion of LINE from the intergenic region. The deletion is restricted to benthics and freshwater stickleback populations from British Columbia and Alaska, suggesting it originated when marine stickleback colonized freshwater habitats in this region (**Fig. 5.2**).

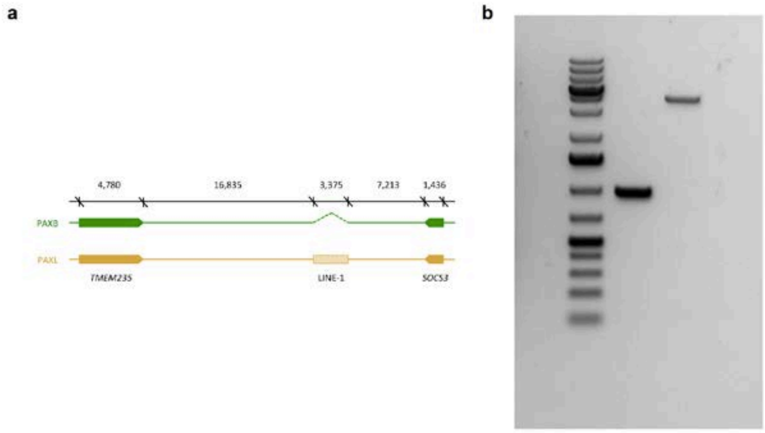

Figure 5.1 | Deletion of long interspersed nuclear element (LINE) in the intergenic region of *SOCS3*. **a**, There is a deletion ~7 kb downstream of *SOCS3* gene in Paxton Lake benthics (PAXB) but not in limnetics (PAXL). The deletion removes a LINE-1 retrotransposon from the region. The sizes of genes and deletion were plotted on top of the gene model. **b,** The deletion in PAXB is confirmed by PCR amplification of the region. **Note: PCR amplication was performed by Ms. Li Ying Tan.**

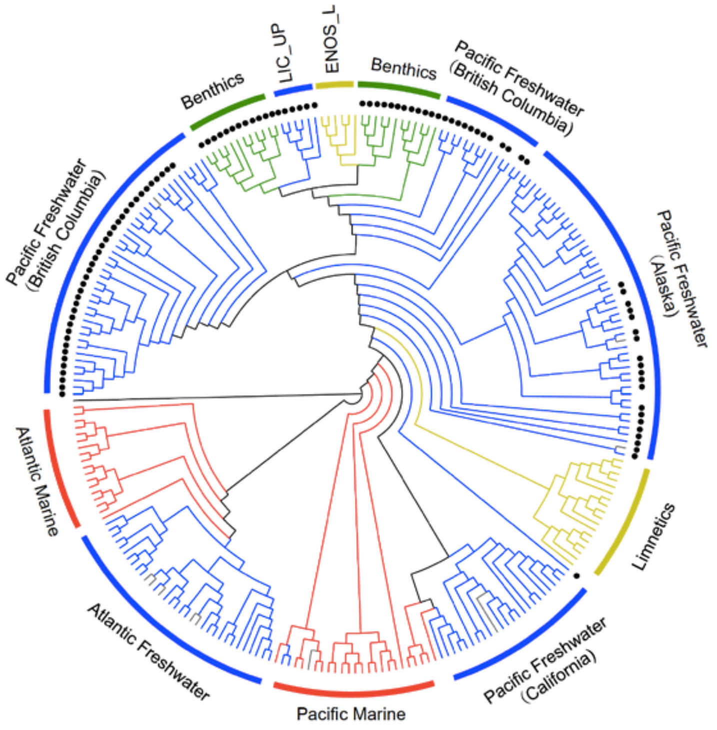
Benthics
LIC_UP
ENOS_L
Benthics
Pacific Freshwater (British Columbia)
Pacific Freshwater (Alaska)
Limnetics
Pacific Freshwater (California)
Pacific Marine
Atlantic Freshwater
Atlantic Marine
Pacific Freshwater (British Columbia)

Figure 5.2 | The deletion in intergenic region of *SOCS3* originated in the region of British Columbia and Alaska. The presence and absence of the deletion were annotated on the maximum-likelihood (ML) tree of benthics and limnetics as well as global marine and freshwater sticklebacks. The presence of deletion in an individual was denoted as black dot on the tree. The deletion is only presented in benthics and freshwater sticklebacks from British Columba and Alaska.

SOCS3 is one of the adaptive *cis*-regulatory diverging genes of Paxton Lake benthics and limnetics. In addition, *SOCS3* regulates lateral line neuromast development in zebrafish. It indicates that the divergence in *cis*-regulatory element of *SOCS3* may contribute to adaptive morphological divergence between Paxton Lake benthics and limnetics. Thus, I collaborated with my colleague Ms. Li Ying Tan to investigate the biological functions of the downstream intergenic region (chrXI: 9,048,002-9,065,075) of *SOCS3* using green fluorescent protein (GFP) reporter assay. As the region of interest is large (~17kb), the reporter constructs were constructed using a recombineering-based approach with bacterial artificial chromosome (BAC). As the BAC libraries were just constructed for Paxton Lake benthics and marine sticklebacks from Salmon River, Alaska and Paxton Lake limnetics carry marine haplotype at the intergenic region of *SOCS3*, the reporter assay was performed for intergenic regions from Paxton Lake benthics and marine sticklebacks. The reporter assay showed there was a clear divergence between the activities of enhancers of *SOCS3* from benthics and marine sticklebacks from Salmon River, Alaska (SALR) (**Fig. 5.3**). Only the enhancer of marine sticklebacks but not benthics drove GFP expression in the pigmentation cells. This suggests the divergence in the enhancers of *SOCS3* contribute to pigmentation divergence between Paxton Lake benthics and limnetics. Benthic and limnetic fish differ in their pigmentation patters (benthics are more melanized) while limnetics have a high degree of silver counter shading (McPhail 1994) (Fig. 5.4). Further there is some evidence that female benthics and limnetics distinguish conspecific males according to body color (Boughman et al 2005). It is therefore possible that *cis*-regulation of *SOCS3* might be subject to natural and/or sexual selection of benthics and limnetics by regulating skin pigmentation.

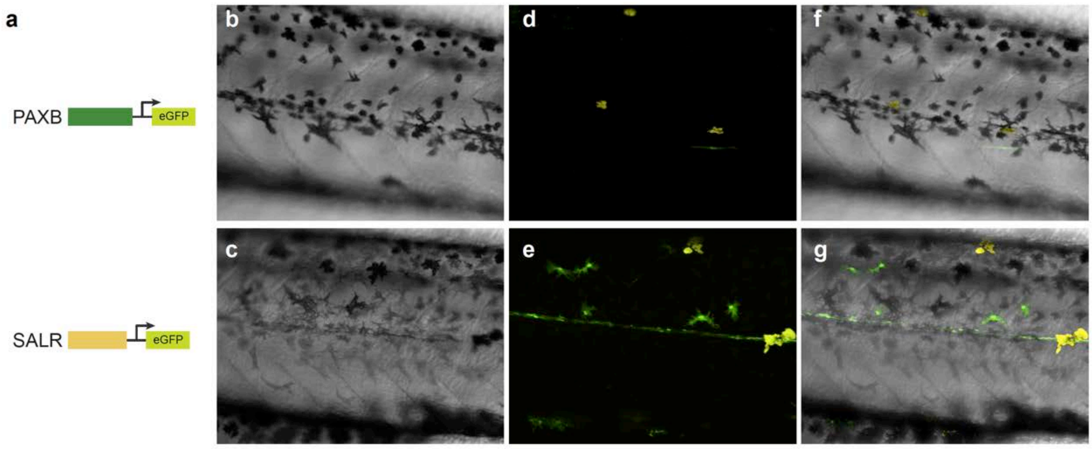

Figure 5.3 | Functional test of enhancer of *SOCS3*. Green fluorescent protein (GFP) reporter essay was performed for enhancer of *SOCS3*. **a**, reporter constructs. **b-c** Bright field images. **d**, Enhancer of *SOCS3* from Paxton Lake benthics does not drives EGFP (green) expression in pigmentation cells. **e**, Enhancer of *SCOS3* from marine population (Salmon River, SALR) drives EGFP (green) expression in pigmentation cells. **e-f**, composite images of corresponding EGFP essay.

Note: Enhancer essay of *SOCS3* is performed by my colleague, Ms. Li Ying Tan.

The collagen family is the one of the most important structure protein families and regulates a variety of developmental processes (Ricard-Blum 2011). Collagens regulate the proliferation and differentiation of cell and therefore control the organization and shape of tissues. The analysis of adaptive regions of benthics and limnetics found two collagen genes (*COL24A1*, *COL7A1*) contribute to the species adaptation. In addition, GO enrichment analysis using human orthologues shown significant enrichment of genes involved in collagen fibril organization. Therefore, to better understand the function of collagen genes in the adaptation of benthics and limnetics. I evaluated the CSS at collagen genes of cross lake benthics and limnetics. There are three collagen genes (*COL21A1*, *COL14A1B*, *COL7A1*) have extreme CSS scores of cross lake benthics and limnetics (top 0.5%) (**Appendix Table 15**).

COL21A1 (chrVI: 7,710,406-7,724,080) has the highest CSS score in the collagen family (**Appendix Table 15**), and two SNPs in the intergenic region have significant nS_L score (FDR<5%) in Paxton Lake benthics. This suggests *COL21A1* was selected in benthics and diverged between benthics and limnetics. Additionally, *COL21A1* showed ASE in three F1 individuals, indicating there is divergence in a *cis*-regulatory element controlling expression of this gene. Thus, functions of the upstream intergenic region (chrVI: 7,700,683-7,724,077) of *COL21A1* were investigated by green fluorescent protein (GFP) reporter assay. The report assay showed the enhancer in the intergenic region of *COL21A1* drove GFP expression in the pigmentation cells (melanophore and xanthophore) (**Fig. 5.4**) It is therefore possible that *cis*-regulation of *COL21A1* might be also subject to natural and/or sexual selection of benthics and limnetics by regulating skin pigmentation.

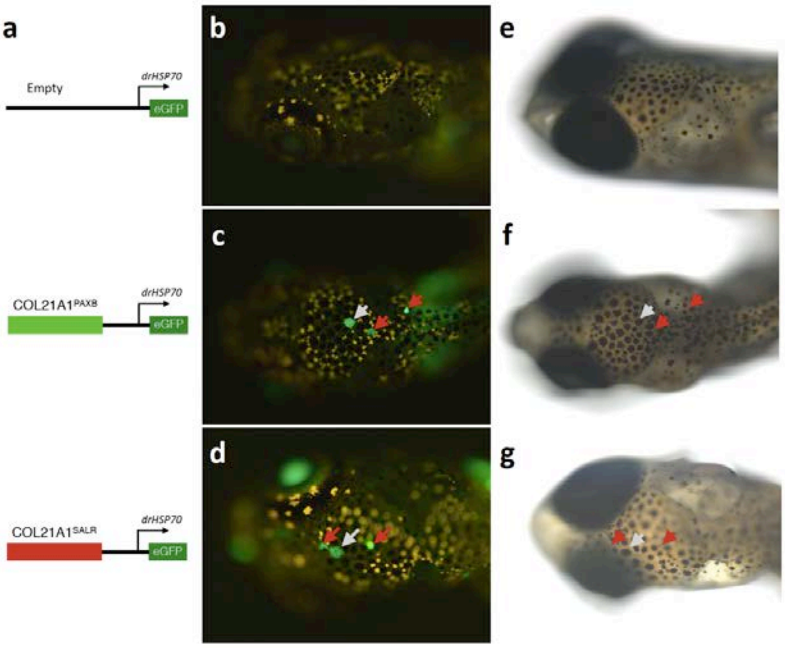

Figure 5.4 | Functional test of enhancer of *COL21A1*. Green fluorescent protein (GFP) reporter essay was performed for enhancer of *COL21A1*. **a**, reporter constructs. **b** negative control. **c**, Enhancer of *COL21A1* from Paxton Lake benthics drives EGFP (green) expression in pigmentation cells (melanophores and xanthophores). White arrows indicate fluorescent signals at melanophores. Red arrows indicate fluorescent signals at xanthophores. **d**, Enhancer of *COL21A1* from marine population (Salmon River, SALR) drives EGFP (green) expression in pigmentation cells (melanophores and xanthophores). **e-g**, Bright field images of corresponding EGFP essay.

Note: Enhancer essay of *COL21A1* is performed by my colleague, Ms. Li Ying Tan.

5.4 Discussion

It has been proposed that genetic changes in regulatory sequences plays an important role in the phenotypic adaptation and evolution (King & Wilson 1975). Recent genomic studies in human and mouse showed local adaptation was largely due to changes in gene expression rather than in coding sequence (Fraser 2011, Fraser 2013). *Cis*-regulatory change is critical for morphological adaptation, as it can modify the morphology of individuals without a cost imposed by more pleiotropic changes in protein structure (Stern & Orgogozo 2008). *Cis*-regulatory is also important for individual's changes responding to environmental changes (Lopez-Maury et al 2008).

Regulatory changes play an important role in the adaptation of marine and freshwater sticklebacks. Genetic studies of stickleback adaptation revealed divergence in several important adaptive morphological traits between marine and freshwater stickleback populations attribute to changes in regulatory sequence (Chan et al 2010, Cleves et al 2014, Miller et al 2007, O'Brown et al 2015). In addition, genomic study of marine and freshwater stickleback adaptation showed most of the adaptive sequence changes located in regulatory sequences. As parallel morphological divergence is observed between benthics and limnetics from different lakes (McPhail 1994), it is likely that *cis*-regulatory changes contribute to the adaptation of these two species. To investigate the role of regulatory changes in benthics and limnetics adaptation, I performed ASE analysis using multiple F1 crosses of wild-caught Paxton Lake benthic and limnetic ecotypes. My analysis shows as much as 10% of genes in the genome have allele specific expression, suggesting *cis*-regulatory changes are of importance to the adaptive divergence of benthics and limnetics. *Cis*-regulatory diverging genes showed significantly enriched in biological processes of otolith development, heart development, ion transport, and organ morphogenesis. In addition, several *cis*-regulatory diverging genes regulating heart development, otolith development, and organ morphogenesis have been subject to divergent selection in benthics and limnetics. Most of these genes have important functions in fish development, and changes in coding sequence of these genes may have functional constraint. Therefore, genetic changes at these genes are most likely through changes in regulatory sequences.

Several *cis*-regulatory diverging genes are highly diverged between benthics and limnetics at regulatory regions, indicating expression divergence at these genes are critical for benthics and limnetics adaptation. Therefore, I collaborated with my colleague, Li Ying Tan, to functional dissect two of these genes (*SOCS3* and *COL21A1*). Interestingly, enhancer reporter assay identified enhancer activities in pigmentation cells for the intergenic regions of both genes. In addition, there is a clear divergence of activities between the *SOCS3* enhancers from Paxton Lake benthics and limnetics. This suggests that *cis*-regulatory divergence of *SOCS3* contribute to the pigmentation divergence between Paxton Lake benthics and limnetics. Divergence in the intergenic regions of *COL21A1* also contributes to the pigmentation divergence, possibly through incorporating divergence in *trans*-acting factors.

5.5 Methods

5.5.1 Sequencing and SNP calling of parental fishes

5.5.1.1 Sample processing and sequencing (Note: this step was performed by Dr. Jukka-Pekka Verta)

Genomic DNA of parental fishes of each F1 crosses was extracted from fin samples following the protocol described previously (Peichel et al 2001). Due to the yield of DNA from tiny fin chips, DNA sequencing libraries were constructed using Tn5 transposase expressed in-house as previously described (Picelli et al 2014). Genomic DNA was purified using AmpureXP bead (Beckman Coulter GmbH, Krefeld, Germany) and “tragmented” by Tn5-transposase. Each tagmented DNA sample was then PCR amplified with Q5 High-Fidelity DNA polymerase (New England Biolabs) using barcoded i7- and i5-index primers. Six parental fishes were pooled and sequenced on one lane of Illumina HiSeq 3000 with 2x150 bp chemistry at the Genome Core Facility at the Max Plank Institute for Developmental Biology.

5.5.1.2 SNP calling and filtering

DNA-sequencing reads were aligned to stickleback gasAcu1 reference sequence using BWA v0.7.10-r789 with BWA `mem` function. The SNPs of parental fishes were identified following the SNP calling pipeline described in **Section 6.1.3** using GATK v3.4. As GATK HaplotypeCaller improves SNP calling quality by constructing correlation matrix of multiple samples, increasing the number of samples used in SNP calling step using HaplotypeCaller is recommended. Therefore, SNP calling was performed for all 8 parental individuals simultaneously. Raw SNPs were filtered using VQSR function of GATK. Due to the lack of “golden” quality reference variant set for sticklebacks, I generated training variant set used in VQSR by hard-filtering the raw variant calls of 8 parental individuals with parameters “QD < 2.00 || FS > 60.000 || MQ < 50.00 || MQRankSum < -12.500 || ReadPosRankSum < -8.000”. SNPs were filtered with 99.9% sensitivity tranche to retain maximum number of SNP in the dataset.

5.5.2 RNA-sequencing and data processing

5.5.2.1 Sample processing and sequencing (Note: this step was performed by Dr. Jukka-Pekka Verta)

mRNA was extracted using whole fish of F1 individual two months after fertilization. Strand-specific RNA-seq libraries were constructed using TruSeq Stranded RNA-seq kit with modified protocol. The insert size of sequencing library was optimized to center ~290 bp. RNA-seq libraries of 16 F1 individuals were pooled and sequenced on one lane of Illumina HiSeq 3000 with 2x150 bp chemistry at the Genome Core Facility at the Max Plank Institute for Developmental Biology.

5.5.2.2 RNA-seq reads alignment and processing

RNA-seq reads were trimmed for low-quality ends of reads and adapter sequencing using Trim Galore program (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) with parameter “--illumina --stringency 5 --quality 20 --pair”. Trimmed RNA-seq reads were aligned to stickleback gasAcu1 reference sequence with multisample two-pass mode of STAR aligner (Dobin et al 2013) using Ensembl stickleback gene model v90 as guidance. First, RNA-seq reads of each F1 individual were aligned to reference sequence with parameters:

“--outFilterIntronMotif RemoveNoncanonicalUnannotated --chimSegmentMin 50 --outFilterMultimapNmax 10 --alignSJDBoverhandMin 1 --alignIntronMin 20 --alignIntronMax 200000 --quantMode GeneCounts”

Second, RNA-seq reads of each F1 individuals were aligned to reference sequence again with the guidance of alignments of all 16 F1 individuals generated in previous step with parameters:

“--outFilterIntronMotifs RemoveNoncanonicalUnannotated --chimSegmentMin 50 --outFilterMultimapNmax 10 --outFilterType BySJout --alignSJDBoverhangMin 1 --alignIntronMin 20 --alignIntronMax 200000 --alignMatesGapMax 200000 --quantMode GeneCounts --limitSjdbInsertNsj 1500000”

I assembled the aligned RNA-seq reads of each F1 individual into transcripts using cufflinks v2.2.1 (Trapnell et al 2012). Transcript assembly was guided using Ensembl stickleback gene model v90 with parameters:

“`--min-intron-length 20 --library-type fr-firststrand --multi-read-correct --min-isoform-fraction 0.15 --min-frags-per-transfrag 20 --max-multiread-fraction 0.5`”.

The assembled transcripts of individuals from the same F1 cross were then merged as a single transcriptome assembly using cuffmerge program of cufflinks package. In addition, a single transcriptome assembly of all 16 F1 individuals was generated and used in the following ASE analysis. The transcriptome assemblies of each F1 cross and all individuals were summarized and compared to Ensembl gene model v90 using cuffcompare program of cufflinks package.

5.5.3 Allele-specific expression (ASE) analysis

High-confidence informative SNP set of parental fishes was generated for ASE analysis. The high-confidence informative SNPs were defined with two criteria: first, parental fishes of each F1 cross are homozygous for different alleles at the SNP site; second, the genotype calls of both alleles at the SNP site are supported by at least 10 sequencing reads. To avoid mapping bias of RNA-seq reads at informative SNP sites, I used the FastaAlternateReferenceMaker function of GATK v3.4 to mask the stickleback reference sequence with “N” in the corresponding position. RNA-seq reads were aligned to the “N” masked reference sequence with multisample two-pass mode of STAR aligner using the protocol described previously (**see Section 6.5.2**).

I evaluated the allele-specific expression of F1 individuals as differential read counts overlapping informative SNP sites using ASEReadCounter function of GATK v3.4. To remove the effect of variable sequencing coverage, the read counts of each individual were normalized to total library size of all 4 individuals from one F1 cross with custom R script. ASE at each informative SNP site was test using binomial exact test with an FDR level of 10% with custom R script. SNP sites having allele-specific expression were assigned to transcriptome assembly of all F1 individuals in R with GenomicRanges package. Genes with at least one ASE SNP and one SNP with same direction of differential expression between benthics and limnetics in all 4 individuals from a F1 cross were identified as genes with *cis*-regulatory divergence between benthics and limnetics. Genes with *cis*-regulatory divergence in F1 crosses with same mating direction were combined and used in following analyses.

5.5.4 GO enrichment analysis

GO enrichment analyses of genes with *cis*-regulatory divergence were performed using method described previously (**see Section 8.2.9**). In total, 491 and 559 genes having ASE in reciprocal F1 crosses (benthics x limnetics and limnetics x benthics) have 1-to-1 orthologs in zebrafish separately. GO enrichment analyses were performed using zebrafish orthologs in R with topGO package. GO categories with *P*-value less than 0.05 in the enrichment analyses were retained.

5.5.5 Green fluorescent protein (GFP) reporter essay (Note: All the experiments were performed by Ms. Li Ying Tan)

5.5.5.1 Reporter constructs

Divergent genomic regions were PCR amplified from end-sequenced BAC clones (CHORI, Children's Hospital Oakland Research Institute) spanning the regions of interest (**Table 6.2**). The fragments were then cloned directionally into the reporter plasmid ipCM001 upstream of an eGFP gene fused to a zebrafish minimal *Hsp70* promoter. Minimal *Tol2* recognition sites flank the entire reporter cassette, which allows for the reliable integration of the cassette into the stickleback genome via a "cut-and-paste" mechanism (Urasaki et al 2006).

Table 6.2 Information of reporter assay constructs of studied divergent regions

Coordinates of Divergent Region	Size (bp)	Benthic Allele	Limnetic or Ancestral Allele	Studied Gene
ChrVI: 7,700,683 - 7,724,077	~ 23,400	CHORI-215-44M13	CHORI-213-200K09	*COL21A1*
ChrXI: 9,048,002 – 9,065,075	~ 17,073	CHORI-215-19O12	CHORI-213-193F02	*SOCS3*

The reporter constructs were constructed using a recombineering-based approach. Firstly, end-sequenced BAC clones containing the region of interest from a benthic library (CHORI-215, Paxton Lake) and a marine library (CHORI-213, Salmon River) were electroporated separately into MW005 cells to serve as substrates for recombineering (Westenberg et al 2010). Next, a gene fragment was designed to contain ~150 bp homology arms matching invariant regions flanking the region of interest (Integrated DNA Technologies, USA). The gene fragment was cloned directionally into ipCM001 upstream of the minimal *Hsp70* promoter. The entire plasmid was then linearised and electroporated into the BAC-containing cells. Recombination was induced as described by (Sharan et al 2009)) and subsequent clones were screened for correct homologous recombination by PCR of the left and right junctions.

5.5.5.2 Stickleback transgenics

Transposase mRNA was transcribed from the pCS-TP plasmid as described in (Kawakami et al 2004)). The reporter constructs were co-injected with *Tol2* transposase mRNA into fertilised stickleback embryos at the one-cell stage. The injections were performed at a DNA concentration of 20 ng/µl and an mRNA concentration of 50 ng/µl. The embryos were monitored over their development and screened for positive eGFP expression.

6 GENOMIC BASIS OF REVERSE SPECIATION OF ENOS LAKE BENTHICS AND LIMNETICS

6.1 Background and Aims

Sympatric benthic and limnetic stickleback ecotype pair in Enos Lake was first described as morphologically divergent in 1984 (McPhail 1984). Study in 1992 found the majority of wild caught sticklebacks from Enos Lake were morphologically divergent and about 1% of stickleback individuals collected in the lake were considered as possible hybrids between the two species due to intermediate phenotype (Schluter & McPhail 1992). Later study in 2001 showed that about 12% of sticklebacks collected in Enos Lake have intermediate morphologies between benthics and limnetics, suggesting the species pair in Enos Lake may “collapse” into a hybrid population due to increased hybridization (Kraak et al 2001). By analyzing the morphology of Enos Lake sticklebacks collected from 1977 to 2002, researcher found the increased hybridization might occur between 1994 and 1997 due to the introduction of crayfish in early 1990s (Taylor et al 2006). Both morphological and genetic studies indicated the reverse speciation is a result of introgression from benthics to limnetics (Gow et al 2006, Rudman & Schluter 2016).

During the process of collapse into a hybrid swarm it is anticipated that different parts of the genome show differing degrees and rates of homogenization. The specific loci that have homogenized and those that remain distinct have the potential to offer insight into the genetic basis of speciation. It can be argued that loci that remain distinct between benthics and limnetics despite increased hybridization may be 1) located in genomic regions of low recombination that are more robust to the homogenizing effects of recombination, 2) played a particularly important role in reproductive isolation between the species pairs such that homogenization at these loci still has deleterious fitness effects. In addition, genomic loci that are divergent in other benthic-limnetic species pairs but have homogenized in Enos Lake can inform us about the types of selection pressures relevant to divergent benthic-limnetic adaptation that have changed or been lost in the last 30 years in

Enos Lake. In this chapter, I studied the reverse speciation of Enos Lake benthics and limnetics using whole genome resequencing data. The aims of this chapter are:

- to investigate the pattern of genomic homogenization of Enos Lake benthics and limnetics.
- to determine the biological function of "collapsed" regions in the genome of Enos Lake benthics and limnetics.

6.2 Genomic pattern of reverse speciation of Enos Lake benthics and limnetics

Since Enos Lake fish are now morphologically intermediate, the divergent loci that have since been homogenized in the genome of Enos Lake benthics and limnetics during reverse speciation may play a critical role in maintaining the phenotypic divergence between benthics and limnetics. However, the extent of homogenization between the genome of Enos Lake species pair is unclear. To quantify the extent of genome homogenization during the reverse speciation of Enos Lake benthics and limnetics, I compared the proportion of divergent regions in benthics and limnetics from Enos Lake and other non-collapsed lakes. The genome-wide genetic divergence (F_{ST}) was calculated in 43,926 non-overlapping genomic windows (window size: 10kb) for benthics and limnetics from each lake. The proportion of divergent genomic regions ($F_{ST} > 0.5$) decreased in Enos Lake benthics and limnetics compared to species pairs from other lakes (**Table 6.1**). For example, the proportion of divergent regions reduced from 16.25% in the genomes of Paxton Lake benthics and limnetics to 4.84% in the genomes of Enos Lake pair. There are about 6% of genomic regions showed parallel divergence in the pair-wise comparison of species pairs from lakes (Paxton Lake, Priest Lake, Little Quarry Lake) in which the reverse speciation did not occur (non-collapsed lakes). Only about 1.5% of the genome regions that showed parallel benthic-limnetic divergence between two non-collapsed lakes are also diverged in species pair from Enos Lake. Finally, 4% of the genome showed parallel divergence among the species pairs from all three non-collapsed lakes. Only one fourth of these regions diverged between Enos

Lake benthics and limnetics. This suggests a large portion of divergent regions have been collapsed during reverse speciation of Enos Lake benthics and limnetics.

Table 6.1. The proportion of "collapsed" genomic regions of Enos Lake benthics and limnetics

Lake	No. of windows	Proportion	Lake	No. of windows	Proportion
One Lake					
PAX	7,140	16.25%	PAX+ENS	1,292	2.94%
PRI	5,710	13.00%	PRI+ENS	1,029	2.34%
QRY	4,180	9.52%	QRY+ENS	857	1.95%
ENS	2,125	4.84%			
Two Lakes					
PAX+PRI	3,554	8.09%	PAX+PRI+ENS	690	1.57%
PAX+QRY	2,662	6.06%	PAX+QRY+ENS	706	1.61%
PRI+QRY	2,338	5.32%	PRI+QRY+ENS	477	1.09%
Three Lakes					
PAX+PRI+QRY	1,758	4%	PAX+PRI+QRY+ENS	413	0.94%

To investigate the distribution of homogenized regions in the genomes of Enos Lake benthics and limnetics, I compared the CSS scores of benthics and limnetic from Enos Lake and three other lakes. The homogenization of genome occurred across the whole genome of Enos Lake benthics and limnetics (**Fig. 6.1**). Interestingly, there is a large region on chromosome I has larger CSS scores in benthics and limnetics from Enos Lake than the species pairs from three other lakes (**Fig. 6.1**). This region is one of the chromosome inversions (chrI: 15,472,665-16,811,878) previously identified between Paxton Lake benthics and limnetics (Chan 2009). Investigating the genotypes of cross-lake benthics and limnetics in this region showed benthics and limnetics carried different genotypes of the inversion. The divergence of inversion is fixed in benthics and limnetics from Priest and Enos, and segregates in Paxton and Little Quarry Lakes but the ecotypes are not fixed for alternate alleles (**Fig. 6.2**).

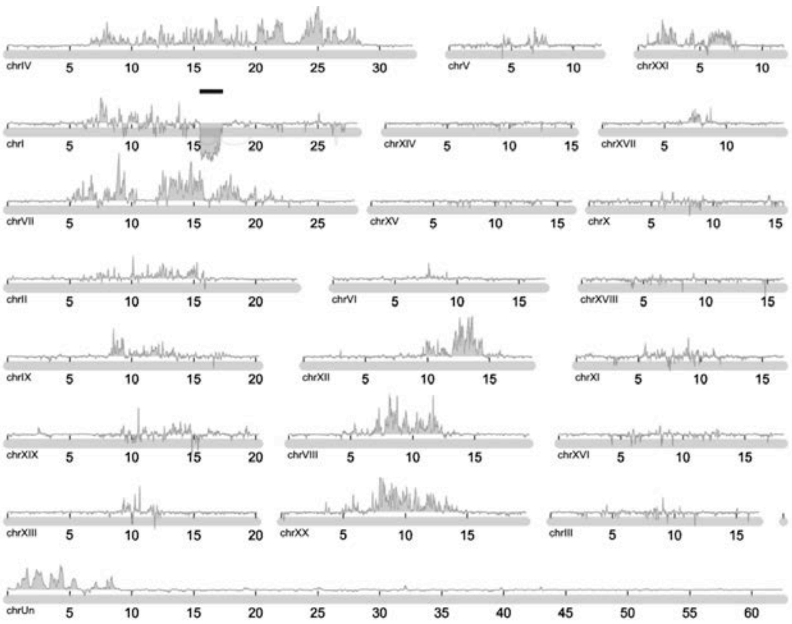

Figure 6.1 | Genomic pattern of CSS difference between Enos and non-collapsed lakes benthic-limnetic species pair. Positive values indicate higher CSS in benthics and limnetics from non-collapsed lakes. Negative values indicate higher CSS in Enos Lake benthics and limnetics. The inversion on chromosome I that is diverged in Enos Lake but not non-collapsed lakes is indicated as black bar on top of the chromosome.

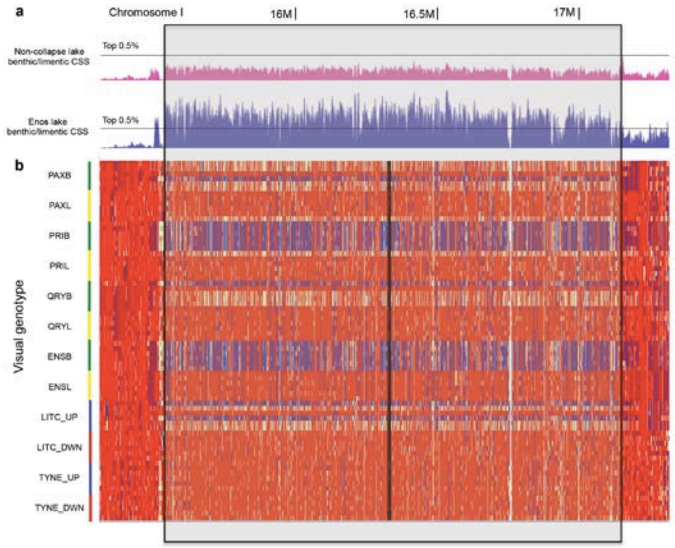

Figure 6.2 | The chromosome I inversion is diverged in Enos Lake Benthics and limnetics. a, CSS scores of non-collapsed and Enos Lake benthics and limnetics. The top 0.5% of genome-wide CSS score is indicated by line. **b,** Visual genotype for benthics and limnetics as well as marine and freshwater ecotypes from Little Campbell River and River Tyne. Red box represents most frequent allele in marine ecotype from Little Campbell River and River Tyne (ancestral allele), blue box represents the alternative allele (derived allele), and yellow box represents heterozygous allele. The chromosome inversion previously identified in Paxton Lake benthics and limnetics is showed as vertical shaded box (Chan 2009).

6.3 Biological functions of “collapsed” regions in Enos Lake benthics and limnetics

The parallel divergent regions in the genomes of non-collapsed lake benthics and limnetics that are homogenized in Enos Lake benthics and limnetics are likely to be particularly important in the reproductive isolation of benthic and limnetic species. Investigating these regions provide valuable insights of benthic and limnetic speciation. Therefore, I studied the functions of genes located in the regions that have the largest difference (top 1%) between CSS of benthics and limnetics from non-collapse lakes and Enos Lake. GO enrichment analysis using human orthologs showed significant enrichment of genes involved in the biological processes of ion transport, muscle development, heart development, lipid localization, regulation of behavior, and response to external stimulus (**Table 6.2**). GO enrichment analysis using zebrafish orthologs showed significant enrichment of genes involved in lipid transport, fluid transport, ion transport, blood vessel development, and signal transduction (**Table 6.3**). It is noteworthy that genes involved in ion transport, muscle development, vascular system development, lipid metabolism, and signal transduction were also enriched in the GO enrichment analysis of genes located in “composite adaptive regions” of benthics and limnetics (**see Section 3.3.2**), emphasizing the importance of these biological processes to the adaptation of benthics and limnetics.

Table 6.2 Enrichment of Gene Ontology categories of human of genes in Enos Lake collapsed regions.

GO category	Annotated	Observed	Expected	*P*-value	Genes included
activation of CREB, activation of CREB transcription factor, CREB activator	11	3	0.29	0.0027	OPRD1B, CAMK1DA, RPS6KA4
regulation of muscle system process	134	10	3.59	0.0032	CAV3, OXTR, COL14A1B, PBE4BA, PTGS2B, TBXA2R, TNNT2A, JUPA, MTMR4
ion transport	1035	42	27.71	0.004	CYP27B1, ITPR1A, SLC4A8, IP6K2B, GRM2A, CAV3, ABCA4A, CNGB3.1, PDE4BA, OPRD1B, PLA2G4AB, PTGS2B, KCNN4, SLC46A1, SLC26A10, ATP13A3, STX1A, CACNA2D3 (1 of many), COX7C, SLC16A1B, ATP8A1, ACSL1A, GRIA1A, ARHGEF9B, GABRA2, SLC16A7, SLC26A5, KCNC2, ZDHHC17, CHRNB1, ATP2A3, CA4A, CNIH2, GABRA3, GRIA3A, SLC12A9
cardiac atrium morphogenesis	25	4	0.67	0.0041	TBX5B, TNNT2A, WNT5A, ENSGACG00000002145
sarcomere organization	26	4	0.7	0.0047	CAV3, FHOD3B, TNNT2A, ENSGACG00000002145
positive regulation of behavior	65	6	1.74	0.0077	SGIP1A, WNT5A, CAMK1DA, si:dkey-11f12.2, DSCAMA, HSPB1
positive regulation of response to external stimulus	107	8	2.86	0.0079	CRY27B1, PTGS2B, WNT5A, CAMK1DA, si:dkey-11f12.2, DSCAMA, HSPB1, PAFAH1B2
cyclic nucleotide metabolic process	130	9	3.48	0.0082	GRM2A, PDE4BA, APLP1, AMPD2B, PDE4CB, AIPL1, WNT5A, GC3
lipid localization	205	12	5.49	0.0091	ABCA4A, PLA2G4AB, KCNN4, OSBPL3B, B4GALNT1A, OSBPL5, LIPCA, ATP8A1, ACSL1A, OSBPL8, SPNS3, VPS51

Table 6.3 Enrichment of Gene Ontology categories of zebrafish of genes in Enos Lake collapsed regions.

GO category	Annotated	Observed	Expected	P-value	Genes included
transmembrane transport	438	17	10.53	0.0341	ITPR1A, SLC2A11A, KCNK5B, SLC46A1, SLC26A10, ABCB4, TRPV6, SLC16A1B, AQP10A, si:ch73-335m24.5, GRIA1A, SLC26A5, KCNC2, SPNS3, SLC8A4A, GRIA3A, SLC12A9
lipid transport	51	5	1.23	0.0073	ATP8A1, SPNS3, VPS51, OSBPL5, OSBPL3B
sulfur compound transport	10	2	0.24	0.0228	SLC26A5, SLC26A10
fluid transport	13	2	0.31	0.0377	SLC8A4A, AQP10A
anion transport	108	6	2.6	0.0454	SLC4A8, SLC26A10, SLC16A1B, AQP10A, ATP8A1, SLC26A5
Golgi vesicle transport	27	3	0.65	0.0262	GOSR1, SEC24D, VPS51
termination of G-protein coupled receptor signaling pathway	27	3	0.65	0.0262	RGS14A, AKAP10, RGS19
ionotropic glutamate receptor signaling pathway	14	2	0.34	0.0432	GRIA1A, GRIA3A
extracellular structure organization	17	3	0.41	0.0072	SEC24D, HMCN1, ITGA5
regulation of DNA-templated transcription, elongation	11	2	0.26	0.0274	TCEA2, TCEA3
negative regulation of neuron death	11	2	0.26	0.0274	PSENEN, GRINAB
nuclear-transcribed mRNA catabolic process, nonsense-mediated decay	12	2	0.29	0.0324	PYM1, SMG9
regulation of autophagy	13	2	0.31	0.0377	MTM1, UVRAG
blood vessel endothelial cell migration	13	2	0.31	0.0377	ITGA5, ENSGACG00000011047
oocyte development	13	2	0.31	0.0377	JUPA, PAQR7B
negative regulation of Wnt signaling pathway	32	3	0.77	0.0407	JUPA, FRMD8, ENSGACG00000020335

6.4 Discussion

In ecological speciation, reproductive isolation can evolved from a byproduct of divergent natural selection if the selected adaptive loci is linked with genomic loci contributing to sexual selection, or a direct product if the hybrids suffer low fitness in both parental habitats due to intermediate phenotypes (Schluter 2009). The persistence of reproductive isolation in sympatric species derived from ecological speciation is attributed to the balance of divergent selection and gene flow (Seehausen 2006). Therefore, sympatric species can “collapse” into a hybrid swarm due to increased if the selective pressure changes due to environmental alteration (Seehausen 2006).

Sympatric benthic and limnetic sticklebacks is once one of the best examples of ecological speciation (Seehausen 2006). Enos Lake benthics and limnetics collected from 1980s to early 1990s show clear morphological divergence, including different body size, body shape, male nuptial color (McPhail 1984). A previous study showed the Enos Lake species pair collected in 1977 and 1988 has distinct morphologies, whereas the morphological divergence is unclear for samples collected in 1997 (Taylor et al 2006). Genetic study using microsatellite markers revealed Enos Lake sticklebacks collected in 1994 were genetically divergent, and the authors proposed the reverse speciation occurred between 1994 and 1997, possibly due to the introduction of crayfish in early 1990s (Taylor et al 2006). Dolph Schluter introduced Enos Lake limnetics to the Murdo Frazer Pond in Vancouver between 1988 and 1989 to preserve the stickleback species pair in Enos Lake. The individuals representing Enos Lake limnetics in this study were collected from Murdo Frazer Pond, and therefore originally considered as typical limnetics. My study of genetic relationship of benthics and limnetics using genome-wide genetic variants (see **Section 2.3.1**) showed Enos Lake limnetics were genetically intermediate between benthics and limnetics. This suggests the increased introgressive hybridization of Enos Lake benthics and limnetics started before 1988, even though the Enos Lake stickleback samples collected at this time have clear morphological divergence. My analyses showed although most of regions have been homogenized between

Enos Lake benthics and limnetics, a few genomic regions are still diverged between the species. Taylor et al (2006) may identify the divergence at these regions when they studied the sample collected in 1994.

The genomic regions that are homogenized between benthics and limnetics from Enos Lake but still diverged in between species pair from other lake are important for the maintenance of reproductive isolation. Genes involved in the biological processes ion transport, muscle development, vascular system development, lipid metabolism, and signal transduction were enriched in the GO enrichment analysis of genes located in the genomic regions that were homogenized in Enos Lake benthics and limnetics. Benthics have less hatching success and survival rate in high salinity environment than limnetics, which is probably due to benthics invaded lakes earlier and adapted to freshwater environment longer than limnetics (Kassen et al 1995). The divergence in the genomic regions regulating ion transport in benthics and limnetics might be resulted from the divergent evolutionary history of these two species. Benthics and limnetics had developed different morphological traits to improve the ability of prey capture (Schluter 1995). For example, benthics have greater hypertrophied epaxial musculature and suction capacity than limnetics to catch benthic invertebrates (McGee et al 2013). The direction of gene flow during reverse speciation in Enos Lake is from benthics to limnetics, and the resulting hybrids are able to consume preys of both benthics (invertebrate) and limnetics (zooplankton) (Rudman & Schluter 2016). Thus, the homogenization of genes controlling muscle development is important for the hybrids to consume food of both benthics and limnetics. Lastly, as the oxygen level and temperature are lower in benthic than in limnetic zone of a freshwater lake (Chiras 2013), benthics might need to develop stronger cardiovascular system to survive in low-oxygen and cool environment. Therefore, the homogenization of genes controlling cardiovascular system development could allow the hybrids to explore benthic habitat and consume food of benthics. In all, genes regulating ion transport, muscle and vascular development are critical for the adaptation of benthics and limnetics, homogenization of these genes facilitate the hybrids in Enos Lake to explore both benthic and limnetic habitats.

6.5 Methods

6.5.1 Comparison of genetic divergence between non-collapsed lake and Enos Lake benthics and limnetics

Genetic divergence of both non-collapsed lake (Paxton, Priest, Little Quarry Lake) and Enos Lake benthics and limnetics was estimated using CSS scores. CSS scores were calculated using the method as described previously (**see Section 2.7.5**). To investigate the genome-wide distribution of homogenized regions in Enos Lake benthics and limnetics, the difference of CSS scores between non-collapsed lake and Enos Lake benthics and limnetics were plotted along chromosomes using custom R script.

The genome-wide extent of Enos Lake reverse speciation was estimated by calculating F_{ST} of benthics and limnetics from different lakes separately in non-overlapping windows (size: 10kb). F_{ST} was calculated using VCFtools v0.1.14. The plots were generated using custom R script.

6.5.2 GO enrichment analysis

GO enrichment analysis of genes in the “collapsed” genomic regions of Enos Lake benthics and limnetics was performed using method described previously (**see Section 3.7.2**). In total, 161 and 116 genes have 1-to-1 orthologs in zebrafish and human separately and the corresponding orthologs were used to perform GO enrichment analyses. GO categories with *P*-value less than 0.05 and 0.01 for analyses using zebrafish and human orthologs were retained.

7 Summary and Perspectives

In chapter 2, I investigated the genomic patterns of adaptive divergence between benthics and limnetics. My analysis revealed there was parallel genetic divergence between benthics and limnetics and about ~10% of genome was consistently diverged among species pairs from all four lakes. In addition, my work showed parallel genetic divergence between benthics and limnetics from different lake attribute to strong divergent natural selection but mostly selection in benthics, in which derived and ancestral alleles were selectively favored by benthics and limnetics respectively.

In chapter 3, I studied the sources and functions of adaptive variation in benthics and limnetics. My analysis found the benthics and limnetics largely used standing genetic variations in their adaptation and the divergence between the species pair was mainly mediated by pre-existing adaptive divergence that facilitated the divergence between marine and freshwater sticklebacks from nearby freshwater system. In addition, I identified several genes that contribute to the adaptation of benthics and limnetics. Some of genes regulate important adaptive traits in sticklebacks, including eye development, body development, and epithelium morphogenesis. These genes can be used in future functional dissections. In addition, genes involved in cardiovascular system development and muscle development are also enriched in adaptive regions of benthics and limnetics, suggesting divergence in genes involved in these two biological processes are important for benthics and limnetics adaptation.

In chapter 4, I inferred the demographic model of benthics and limnetics speciation. I found direct evidence that benthics and limnetics were derived from allopatric speciation, in which the ancestors of Paxton Lake benthics and limnetics invaded the lake at 7,000 and 5,000 years ago respectively.

In chapter 5, I investigated the gene expression divergence of Paxton Lake benthics and limnetics. My analysis showed *cis*-regulatory changes plays an important role in their adaptation. In addition, I collaborated with my colleague to functional dissected two *cis*-regulatory diverging genes. Our results showed the *cis*-regulatory divergence at these two genes contribute to the pigmentation divergence between Paxton Lake benthics and limnetics

In chapter 6, I dissected the genetic basis of reverse speciation of Enos Lake benthics and limnetics. I found the reverse speciation of Enos Lake benthics and limnetics started before 1988, which is earlier than the previous prediction. In addition, several highly divergent regions of benthics and limnetics have been homogenized in the genome of Enos Lake benthics and limnetics. Genes located in these regions showed significantly enriched in the biological processes of ion transport, muscle development, vascular system development, lipid metabolism, and signal transduction. This suggests genes involved in these processes are important for the maintenance of reproductive isolation between benthics and limnetics.

In my study, I have provided insights into the genetic basis of benthic and limnetic stickleback adaptation and speciation. There are still several experiments or analyses that I can perform to further our understanding of this process. First, in my allele specific expression analysis, I did not sequence the parental individuals of F1 crosses. Therefore, I cannot investigate gene expression divergence that has a *trans*-regulatory or *cis*+*trans*- regulatory basis. By sequencing the transcriptome of parental individuals, I can dissect the gene expression divergence of benthics and limnetics comprehensively. Second, I found that several adaptive regions of benthics and limnetics located in regulatory regions in my analysis. However, the resolution of my analysis is not high enough. The current development of chromatin immunoprecipitation sequencing (ChIP-Seq) allows researcher to identify and study the enhancer regions with unprecedented high resolution (less than 100bp) (Park 2009). By combining the results of adaptive region identification and ChIP-Seq analysis, I can further increase the resolution of identifying divergent enhancers between benthics and limnetics, which will facilitate future functional dissection experiments.

8 Reference

Alexander C, Votruba M, Pesch UE, Thiselton DL, Mayer S, et al. 2000. OPA1, encoding a dynamin-related GTPase, is mutated in autosomal dominant optic atrophy linked to chromosome 3q28. *Nat Genet* 26: 211-5

Alonso-Blanco C, Andrade J, Becker C, Bemm F, Bergelson J, et al. 2016. 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana. *Cell* 166: 481-91

Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, Chakravarti A, et al. 2015. A global reference for human genetic variation. *Nature* 526: 68-+

Amsterdam A, Nissen RM, Sun ZX, Swindell EC, Farrington S, Hopkins N. 2004. Identification of 315 genes essential for early zebrafish development. *Proceedings of the National Academy of Sciences of the United States of America* 101: 12792-97

Arnegard ME, McGee MD, Matthews B, Marchinko KB, Conte GL, et al. 2014. Genetics of ecological divergence during speciation. *Nature* 511: 307-11

Barboric M, Lenasi T, Chen H, Johansen EB, Guo S, Peterlin BM. 2009. 7SK snRNP/P-TEFb couples transcription elongation with alternative splicing and is essential for vertebrate development. *Proceedings of the National Academy of Sciences of the United States of America* 106: 7798-803

Barluenga M, Stolting KN, Salzburger W, Muschick M, Meyer A. 2006. Sympatric speciation in Nicaraguan crater lake cichlid fish. *Nature* 439: 719-23

Barrett RD, Rogers SM, Schluter D. 2008. Natural selection on a major armor gene in threespine stickleback. *Science* 322: 255-7

Barrett RD, Schluter D. 2008a. Adaptation from standing genetic variation. *Trends in ecology & evolution* 23: 38-44

Barrett RDH, Hoekstra HE. 2011. Molecular spandrels: tests of adaptation at the genetic level. *Nature Reviews Genetics* 12: 767-80

Barrett RDH, Schluter D. 2008b. Adaptation from standing genetic variation. *Trends in ecology & evolution* 23: 38-44

Barton NH. 2000. Genetic hitchhiking. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 355: 1553-62

Barton NH. 2010. Genetic linkage and natural selection. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 365: 2559-69

Bateson W. 1909. Heredity and Variation in Modern Lights In *Darwin and Modern Science*, ed. AC Seward. Cambridge: Cambridge University Press

Bell M, Foster S. 1994a. *The evolutionary biology of the threespine stickleback*. Oxford: Oxford University Press.

Bell MA, Foster SA. 1994b. Introduction to the evolutionary biology of the threespine stickleback In *The evolutionary biology of the threespine stickleback*, ed. MA Bell, SA Foster. Oxford: Oxford University Press

Berner D, Salzburger W. 2015. The genomics of organismal diversification illuminated by adaptive radiations. *Trends in Genetics* 31: 491-99

Beunders G, Voorhoeve E, Golzio C, Pardo LM, Rosenfeld JA, et al. 2013. Exonic Deletions in AUTS2 Cause a Syndromic Form of Intellectual Disability and Suggest a Critical Role for the C Terminus. *American journal of human genetics* 92: 210-20

Bohne A, Brunet F, Galiana-Arnoux D, Schultheis C, Volff JN. 2008. Transposable elements as drivers of genomic and biological diversity in vertebrates. *Chromosome Res* 16: 203-15

Bolnick DI, Fitzpatrick BM. 2007. Sympatric speciation: Models and empirical evidence. *Annu Rev Ecol Evol S* 38: 459-87

Boughman JW. 2001. Divergent sexual selection enhances reproductive isolation in sticklebacks. *Nature* 411: 944-8

Boughman JW, Rundle HD, Schluter D. 2005. Parallel evolution of sexual isolation in sticklebacks. *Evolution; international journal of organic evolution* 59: 361-73

Brawand D, Wagner CE, Li YI, Malinsky M, Keller I, et al. 2014. The genomic substrate for adaptive radiation in African cichlid fish. *Nature* 513: 375-81

Burri R, Nater A, Kawakami T, Mugal CF, Olason PI, et al. 2015. Linked selection and recombination rate variation drive the evolution of the genomic landscape of differentiation across the speciation continuum of Ficedula flycatchers. *Genome Res* 25: 1656-65

Camus S, Quevedo C, Menendez S, Paramonov I, Stouten PFW, et al. 2012. Identification of phosphorylase kinase as a novel therapeutic target through high-throughput screening for anti-angiogenesis compounds in zebrafish. *Oncogene* 31: 4333-42

Carney TJ, Feitosa NM, Sonntag C, Slanchev K, Kluger J, et al. 2010. Genetic Analysis of Fin Development in Zebrafish Identifies Furin and Hemicentin 1 as Potential Novel Fraser Syndrome Disease Genes. *PLoS genetics* 6

Carroll SB. 2008. Evo-devo and an expanding evolutionary synthesis: A genetic theory of morphological evolution. *Cell* 134: 25-36

Cech JN, Peichel CL. 2015. Identification of the centromeric repeat in the threespine stickleback fish (Gasterosteus aculeatus). *Chromosome Res* 23: 767-79

Chan YF. 2009. *The genomic basis of parallel evolution in three-spined stickleback (gasterosterus aculeatus)*. Stanford University. 194 pp.

Chan YF, Marks ME, Jones FC, Villarreal G, Jr., Shapiro MD, et al. 2010. Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a Pitx1 enhancer. *Science* 327: 302-5

Charlesworth B. 2009. Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation. *Nature reviews. Genetics* 10: 195-205

Charlesworth B, Charlesworth D. 2017. Population genetics from 1966 to 2016. *Heredity* 118: 2-9

Charlesworth B, Morgan MT, Charlesworth D. 1993. The effect of deleterious mutations on neutral molecular variation. *Genetics* 134: 1289-303

Chen H, Patterson N, Reich D. 2010. Population differentiation as a test for selective sweeps. *Genome Res* 20: 393-402

Chen X, Gays D, Milia C, Santoro MM. 2017. Cilia Control Vascular Mural Cell Recruitment in Vertebrates. *Cell reports* 18: 1033-47

Chiras DD. 2013. *Environmental science*. Burlington, MA: Jones and Bartlett Learning.

Cleves PA, Ellis NA, Jimenez MT, Nunez SM, Schluter D, et al. 2014. Evolved tooth gain in sticklebacks is associated with a cis-regulatory allele of Bmp6.

Proceedings of the National Academy of Sciences of the United States of America 111: 13912-7

Colosimo PF, Hosemann KE, Balabhadra S, Villarreal G, Jr., Dickson M, et al. 2005. Widespread parallel evolution in sticklebacks by repeated fixation of Ectodysplasin alleles. *Science* 307: 1928-33

Colosimo PF, Peichel CL, Nereng K, Blackman BK, Shapiro MD, et al. 2004. The genetic architecture of parallel armor plate reduction in threespine sticklebacks. *PLoS biology* 2: E109

Conte GL, Arnegard ME, Best J, Chan YF, Jones FC, et al. 2015. Extent of QTL Reuse During Repeated Phenotypic Divergence of Sympatric Threespine Stickleback. *Genetics*

Conte GL, Arnegard ME, Peichel CL, Schluter D. 2012. The probability of genetic parallelism and convergence in natural populations. *P Roy Soc B-Biol Sci* 279: 5039-47

Coop G, Pickrell JK, Novembre J, Kudaravalli S, Li J, et al. 2009. The role of geography in human adaptation. *PLoS genetics* 5: e1000500

Coyne JA. 2007. Sympatric speciation. *Current biology : CB* 17: R787-R88

Coyne JA, Orr HA. 2004. *Speciation*. Sunderland: Sinauer Associates.

Cresko WA, Amores A, Wilson C, Murphy J, Currey M, et al. 2004. Parallel genetic basis for repeated evolution of armor loss in Alaskan threespine stickleback populations. *Proceedings of the National Academy of Sciences of the United States of America* 101: 6050-5

Crow KD, Munehara H, Bernardi G. 2010. Sympatric speciation in a genus of marine reef fishes. *Mol Ecol* 19: 2089-105

Cruz A, Lombarte A. 2004. Otolith size and its relationship with colour patterns and sound production. *J Fish Biol* 65: 1512-25

Cuthill IC, Allen WL, Arbuckle K, Caspers B, Chaplin G, et al. 2017. The biology of color. *Science* 357

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, et al. 2011. The variant call format and VCFtools. *Bioinformatics* 27: 2156-8

Dang NN, Murrell DF. 2008. Mutation analysis and characterization of COL7A1 mutations in dystrophic epidermolysis bullosa. *Exp Dermatol* 17: 553-68

Darwin C. 1859. *On the origin of Species*. London: John Murray.

Deagle BE, Jones FC, Chan YGF, Absher DM, Kingsley DM, Reimchen TE. 2012. Population genomics of parallel phenotypic evolution in stickleback across stream-lake ecological transitions. *P Roy Soc B-Biol Sci* 279: 1277-86

DeGiorgio M, Huber CD, Hubisz MJ, Hellmann I, Nielsen R. 2016. SweepFinder2: increased sensitivity, robustness and flexibility. *Bioinformatics* 32: 1895-7

Delaneau O, Howie B, Cox AJ, Zagury JF, Marchini J. 2013. Haplotype Estimation Using Sequencing Reads. *American journal of human genetics* 93: 687-96

Delettre C, Lenaers G, Griffoin JM, Gigarel N, Lorenzo C, et al. 2000. Nuclear gene OPA1, encoding a mitochondrial dynamin-related protein, is mutated in dominant optic atrophy. *Nat Genet* 26: 207-10

Dieckmann U, Doebeli M, Metz J, Tautz D. 2004. Introduction In *Adaptive Speciation*, ed. U Dieckmann, M Doebeli, J Metz, D Tautz, pp. 1-17. Cambridge: Cambridge University Press

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, et al. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29: 15-21

Dobzhansky T. 1936. Studies on Hybrid Sterility. II. Localization of Sterility Factors in Drosophila Pseudoobscura Hybrids. *Genetics* 21: 113-35

Duran I, Csukasi F, Taylor SP, Krakow D, Becerra J, et al. 2015. Collagen duplicate genes of bone and cartilage participate during regeneration of zebrafish fin skeleton. *Gene Expr Patterns* 19: 60-69

Eckfeldt CE, Mendenhall EM, Flynn CM, Wang TF, Pickart MA, et al. 2005. Functional analysis of human hematopoietic stem cell gene expression using zebrafish. *PLoS biology* 3: 1449-58

Elbarbary RA, Lucas BA, Maquat LE. 2016. Retrotransposons as regulators of gene expression. *Science* 351

Ellegren H. 2014. Genome sequencing and population genomics in non-model organisms. *Trends in ecology & evolution* 29: 51-63

Ellegren H, Smeds L, Burri R, Olason PI, Backstrom N, et al. 2012. The genomic landscape of species divergence in Ficedula flycatchers. *Nature* 491: 756-60

Elmer KR, Meyer A. 2011. Adaptation in the age of ecological genomics: insights from parallelism and convergence. *Trends in ecology & evolution* 26: 298-306

Elsaeidi F, Bemben MA, Zhao XF, Goldman D. 2014. Jak/Stat Signaling Stimulates Zebrafish Optic Nerve Regeneration and Overcomes the Inhibitory Actions of Socs3 and Sfpq. *J Neurosci* 34: 2632-44

Erickson PA, Ellis NA, Miller CT. 2016. Microinjection for Transgenesis and Genome Editing in Threespine Sticklebacks. *Journal of Visualized Experiments*

Faria R, Renaut S, Galindo J, Pinho C, Melo-Ferreira J, et al. 2014. Advances in Ecological Speciation: an integrative approach. *Mol Ecol* 23: 513-21

Farooq M, Sulochana KN, Pan XF, To JW, Sheng D, et al. 2008. Histone deacetylase 3 (hdac3) is specifically required for liver development in zebrafish. *Developmental Biology* 317: 336-53

Fay JC, Wu CI. 2000. Hitchhiking under positive Darwinian selection. *Genetics* 155: 1405-13

Feder JL, Egan SP, Nosil P. 2012. The genomics of speciation-with-gene-flow. *Trends in genetics : TIG* 28: 342-50

Felsenstein J. 1981. Skepticism Towards Santa Rosalia, or Why Are There So Few Kinds of Animals. *Evolution; international journal of organic evolution* 35: 124-38

Ferchaud AL, Hansen MM. 2016. The impact of selection, gene flow and demographic history on heterogeneous genomic divergence: three-spine sticklebacks in divergent environments. *Mol Ecol* 25: 238-59

Fernald RD. 1984. Vision and Behavior in an African Cichlid Fish. *Am Sci* 72: 58-65

Ferrer-Admetlla A, Liang M, Korneliussen T, Nielsen R. 2014. On detecting incomplete soft or hard selective sweeps using haplotype structure. *Molecular biology and evolution* 31: 1275-91

Fisher RA. 1958. *The Genetical Theory of Natural Selection*. New York: Dover Publications.

Fraser HB. 2011. Genome-wide approaches to the study of adaptive gene expression evolution Systematic studies of evolutionary adaptations involving gene expression will allow many fundamental questions in evolutionary biology to be addressed. *Bioessays* 33: 469-77

Fraser HB. 2013. Gene expression drives local adaptation in humans. *Genome Res* 23: 1089-96

Fumagalli M, Moltke I, Grarup N, Racimo F, Bjerregaard P, et al. 2015. Greenlandic Inuit show genetic signatures of diet and climate adaptation. *Science* 349: 1343-47

Garcia J, Bagwell J, Njaine B, Norman J, Levic DS, et al. 2017. Sheath Cell Invasion and Trans-differentiation Repair Mechanical Damage Caused by Loss of Caveolae in the Zebrafish Notochord. *Current Biology* 27: 1982-+

Garud NR, Messer PW, Buzbas EO, Petrov DA. 2015. Recent Selective Sweeps in North American Drosophila melanogaster Show Signatures of Soft Sweeps. *PLoS Genet.* 11

Gavrilets S. 2003. Perspective: models of speciation: what have we learned in 40 years? *Evolution; international journal of organic evolution* 57: 2197-215

Ghysen A, Dambly-Chaudiere C. 2004. Development of the zebrafish lateral line. *Curr Opin Neurobiol* 14: 67-73

Gillespie JH. 1991. *The Causes of Molecular Evolution*. Oxford, UK: Oxford University Press.

Gillespie JH. 2004. *Population genetics : a concise guide*. Baltimore, Md.: Johns Hopkins University Press. xiv, 214 p. pp.

Gislason D, Ferguson M, Skulason S, Snorrason SS. 1999. Rapid and coupled phenotypic and genetic divergence in Icelandic Arctic char (Salvelinus alpinus). *Can J Fish Aquat Sci* 56: 2229-34

Goncalves A, Leigh-Brown S, Thybert D, Stefflova K, Turro E, et al. 2012. Extensive compensatory cis-trans regulation in the evolution of mouse gene expression. *Genome Res* 22: 2376-84

Gow JL, Peichel CL, Taylor EB. 2006. Contrasting hybridization rates between sympatric three-spined sticklebacks highlight the fragility of reproductive barriers between evolutionarily young species. *Mol Ecol* 15: 739-52

Graur D, Li W. 2000. *Fundamentals of Molecular Evolution*. Sunderland, MA: Sinauer Associates.

Grossman SR, Andersen KG, Shlyakhter I, Tabrizi S, Winnicki S, et al. 2013. Identifying recent adaptations in large-scale genomic data. *Cell* 152: 703-13

Grossman SR, Shlyakhter I, Karlsson EK, Byrne EH, Morales S, et al. 2010. A composite of multiple signals distinguishes causal variants in regions of positive selection. *Science* 327: 883-6

Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. 2009a. Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data. *PLoS genetics* 5

Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. 2009b. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS genetics* 5: e1000695

Han F, Lamichhaney S, Grant BR, Grant PR, Andersson L, Webster MT. 2017. Gene flow, ancient polymorphism, and ecological adaptation shape the genomic landscape of divergence among Darwin's finches. *Genome Res* 27: 1004-15

Harr B. 2006a. Genomic islands of differentiation between house mouse subspecies. *Genome Res* 16: 730-37

Harr B. 2006b. Genomic islands of differentiation between house mouse subspecies. *Genome Res* 16: 730-37

Harris H. 1966. Enzyme Polymorphisms in Man. *Proc R Soc Ser B-Bio* 164: 298-310

Hart JC, Miller CT. 2017. Sequence-Based Mapping and Genome Editing Reveal Mutations in Stickleback Hps5 Cause Oculocutaneous Albinism and the casper Phenotype. *G3* 7: 3123-31

Hatfield T. 1997. Genetic divergence in adaptive characters between sympatric species of stickleback. *The American naturalist* 149: 1009-29

Hatfield T, Schluter D. 1999. Ecological speciation in sticklebacks: Environment-dependent hybrid fitness. *Evolution; international journal of organic evolution* 53: 866-73

Haubold B, Pfaffelhuber P, Lynch M. 2010. mlRho - a program for estimating the population mutation and recombination rates from shotgun-sequenced diploid genomes. *Mol Ecol* 19 Suppl 1: 277-84

He F, Zhang X, Hu JY, Turck F, Dong X, et al. 2012. Genome-wide Analysis of Cis-regulatory Divergence between Species in the Arabidopsis Genus. *Molecular biology and evolution* 29: 3385-95

He YZ, Wang ZM, Sun SY, Tang DM, Li WY, et al. 2016. HDAC3 Is Required for Posterior Lateral Line Development in Zebrafish. *Mol Neurobiol* 53: 5103-17

Hedrick PW. 2005. *Genetics of populations*. Boston: Jones and Bartlett Publishers. xiii, 737 p. pp.

Hedrick PW. 2013. Adaptive introgression in animals: examples and comparison to new mutation and standing variation as sources of adaptive variation. *Mol Ecol* 22: 4606-18

Hemmer-Hansen J, Nielsen EE, Therkildsen NO, Taylor MI, Ogden R, et al. 2013. A genomic island linked to ecotype divergence in Atlantic cod. *Mol Ecol* 22: 2653-67

Hermisson J, Pennings PS. 2005. Soft sweeps: molecular population genetics of adaptation from standing genetic variation. *Genetics* 169: 2335-52

Hewitt GM. 1988. Hybrid zones-natural laboratories for evolutionary studies. *Trends in ecology & evolution* 3: 158-67

Hoekstra HE, Coyne JA. 2007. The locus of evolution: Evo devo and the genetics of adaptation. *Evolution; international journal of organic evolution* 61: 995-1016

Hoekstra HE, Hirschmann RJ, Bundey RA, Insel PA, Crossland JP. 2006. A single amino acid mutation contributes to adaptive beach mouse color pattern. *Science* 313: 101-04

Hoekstra HE, Nachman MW. 2003. Different genes underlie adaptive melanism in different populations of rock pocket mice. *Mol Ecol* 12: 1185-94

Hohenlohe PA, Bassham S, Etter PD, Stiffler N, Johnson EA, Cresko WA. 2010. Population Genomics of Parallel Adaptation in Threespine Stickleback using Sequenced RAD Tags. *PLoS genetics* 6

Holsinger KE, Weir BS. 2009a. Genetics in geographically structured populations: defining, estimating and interpreting F-ST. *Nature Reviews Genetics* 10: 639-50

Holsinger KE, Weir BS. 2009b. Genetics in geographically structured populations: defining, estimating and interpreting FST. *Nature reviews. Genetics* 10: 639-50

Huang LW, Xiao A, Choi SY, Kan QN, Zhou WB, et al. 2014. Wnt5a Is Necessary for Normal Kidney Development in Zebrafish and Mice. *Nephron Exp Nephrol* 128: 80-88

Hubby JL, Lewontin RC. 1966. A Molecular Approach to Study of Genic Heterozygosity in Natural Populations .I. Number of Alleles at Different Loci in Drosophila Pseudoobscura. *Genetics* 54: 577-94

Huerta-Sanchez E, Jin X, Asan, Bianba Z, Peter BM, et al. 2014. Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. *Nature* 512: 194-7

Huson DH, Richter DC, Rausch C, Dezulian T, Franz M, Rupp R. 2007. Dendroscope: An interactive viewer for large phylogenetic trees. *Bmc Bioinformatics* 8

Indjeian VB, Kingman GA, Jones FC, Guenther CA, Grimwood J, et al. 2016. Evolving New Skeletal Traits by cis-Regulatory Changes in Bone Morphogenetic Proteins. *Cell* 164: 45-56

Ito Y, Kobayashi S, Nakamura N, Miyagi H, Esaki M, et al. 2013. Close association of carbonic anhydrase (CA2a and CA15a), Na+/H+ exchanger (Nhe3b), and ammonia transporter Rhcg1 in zebrafish ionocytes responsible for Na+ uptake. *Front Physiol* 4

Jensen JD. 2014. On the unfounded enthusiasm for soft selective sweeps. *Nat Commun* 5

Jeong S, Rebeiz M, Andolfatto P, Werner T, True J, Carroll SB. 2008. The evolution of gene regulation underlies a morphological difference between two Drosophila sister species. *Cell* 132: 783-93

Johnson JLFA, Hall TE, Dyson JM, Sonntag C, Ayers K, et al. 2012. Scube activity is necessary for Hedgehog signal transduction in vivo. *Developmental Biology* 368: 193-202

Jones FC, Chan YF, Schmutz J, Grimwood J, Brady SD, et al. 2012a. A genome-wide SNP genotyping array reveals patterns of global and repeated species-pair divergence in sticklebacks. *Current biology : CB* 22: 83-90

Jones FC, Grabherr MG, Chan YF, Russell P, Mauceli E, et al. 2012b. The genomic basis of adaptive evolution in threespine sticklebacks. *Nature* 484: 55-61

Jungke P, Hans S, Gupta M, Machate A, Zoller D, Brand M. 2016. Generation of a conditional lima1a allele in zebrafish using the FLEx switch technology. *Genesis* 54: 19-28

Kassen R, Schluter D, McPhail JD. 1995. Evolutionary history of threespine sticklebacks (Gasterosteus spp) in British Columbia: Insights from a physiological clock. *Canadian Journal of Zoology* 73: 2154-58

Kawakami K, Takeda H, Kawakami N, Kobayashi M, Matsuda N, Mishina M. 2004. A transposon-mediated gene trap approach identifies developmentally regulated genes in zebrafish. *Dev Cell* 7: 133-44

Kim Y, Nielsen R. 2004. Linkage disequilibrium as a signature of selective sweeps. *Genetics* 167: 1513-24

Kim Y, Stephan W. 2002. Detecting a local signature of genetic hitchhiking along a recombining chromosome. *Genetics* 160: 765-77

Kimura M. 1968. Evolutionary Rate at the Molecular Level. *Nature* 217: 624

King MC, Wilson AC. 1975. Evolution at 2 Levels in Humans and Chimpanzees. *Science* 188: 107-16

Kingsley DM, Peichel CL. 2007. The molecular genetics of evolutionary change in sticklebacks In *Biology of the Threespine Stickleback*, ed. S Ostlund-Nilsson, I Mayer, FA Huntingford, pp. 41-81. Florida: CRC Press

Kingsley DM, Zhu BL, Osoegawa K, De Jong PJ, Schein J, et al. 2004. New genomic tools for molecular studies of evolutionary change in threespine sticklebacks. *Behaviour* 141: 1331-44

Kirkpatrick M, Barton N. 2006. Chromosome inversions, local adaptation and speciation. *Genetics* 173: 419-34

Klein C, Mikutta J, Krueger J, Scholz K, Brinkmann J, et al. 2011. Neuron navigator 3a regulates liver organogenesis during zebrafish embryogenesis. *Development* 138: 1935-45

Kraak SBM, Mundwiler B, Hart PJB. 2001. Increased number of hybrids between benthic and limnetic three-spined sticklebacks in Enos Lake, Canada; the collapse of a species pair? *J Fish Biol* 58: 1458-64

Lamichhaney S, Berglund J, Almen MS, Maqbool K, Grabherr M, et al. 2015. Evolution of Darwin's finches and their beaks revealed by genome sequencing. *Nature* 518: 371-5

Larson GL. 1976. Social-Behavior and Feeding Ability of 2 Phenotypes of Gasterosteus-Aculeatus in Relation to Their Spatial and Trophic Segregation in a Temperate Lake. *Can J Zool* 54: 107-21

Leonard WJ, O'Shea JJ. 1998. JAKS AND STATS: Biological implications. *Annu Rev Immunol* 16: 293-322

Lewontin RC, Hubby JL. 1966. A Moleuclar Approach to Study of Genic Heterozygosity in Natural Populations .2. Amount of Variation and Degree of Heterozygosity in Natural Populations of Drosophila Pseudoobscura. *Genetics* 54: 595-609

Li H, Durbin R. 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* 26: 589-95

Li J, Qi M, Li CM, Shi D, Zhang DS, et al. 2014a. Tom70 serves as a molecular switch to determine pathological cardiac hypertrophy. *Cell Res* 24: 977-93

Li JY, Li K, Dong XH, Liang D, Zhao QS. 2014b. Ncor1 and Ncor2 Play Essential but Distinct Roles in Zebrafish Primitive Myelopoiesis. *Dev Dynam* 243: 1544-53

Li MJ, Sham PC, Wang JW. 2010. FastPval: a fast and memory efficient program to calculate very low P-values from empirical distribution. *Bioinformatics* 26: 2897-99

Liang J, Wang DM, Renaud G, Wolfsberg TG, Wilson AF, Burgess SM. 2012. The stat3/socs3a Pathway Is a Key Regulator of Hair Cell Regeneration in Zebrafish stat3/socs3a Pathway: Regulator of Hair Cell Regeneration. *J Neurosci* 32: 10662-73

Lin XY, Rinaldo L, Fazly AF, Xu XL. 2007. Depletion of Med10 enhances Wnt and suppresses Nodal signaling during zebrafish embryogenesis. *Developmental Biology* 303: 536-48

Liu S, Lorenzen ED, Fumagalli M, Li B, Harris K, et al. 2014. Population genomics reveal recent speciation and rapid evolutionary adaptation in polar bears. *Cell* 157: 785-94

Loewe L, Charlesworth B. 2007. Background selection in single genes may explain patterns of codon bias. *Genetics* 175: 1381-93

Lombarte A, Cruz A. 2007. Otolith size trends in marine fish communities from different depth strata. *J Fish Biol* 71: 53-76

Lombarte A, Lleonart J. 1993. Otolith Size Changes Related with Body Growth, Habitat Depth and Temperature. *Environmental Biology of Fishes* 37: 297-306

Long Q, Rabanal FA, Meng DZ, Huber CD, Farlow A, et al. 2013. Massive genomic variation and strong selection in Arabidopsis thaliana lines from Sweden. *Nat Genet* 45: 884-U218

Lopez-Maury L, Marguerat S, Bahler J. 2008. Tuning gene expression to changing environments: from rapid responses to evolutionary adaptation. *Nature reviews. Genetics* 9: 583-93

Lowry DB, Modliszewski JL, Wright KM, Wu CA, Willis JH. 2008. The strength and genetic basis of reproductive isolating barriers in flowering plants. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 363: 3009-21

Lynch M, Bobay LM, Catania F, Gout JF, Rho M. 2011. The Repatterning of Eukaryotic Genomes by Random Genetic Drift. *Annu Rev Genom Hum G* 12: 347-66

Maan ME, Seehausen O, Soderberg L, Johnson L, Ripmeester EAP, et al. 2004. Intraspecific sexual selection on a speciation trait, male coloration, in the Lake Victoria cichlid Pundamilia nyererei. *P Roy Soc B-Biol Sci* 271: 2445-52

Mack KL, Nachman MW. 2017. Gene Regulation and Speciation. *Trends in Genetics* 33: 68-80

Maeda K, Kobayashi Y, Udagawa N, Uehara S, Ishihara A, et al. 2012. Wnt5a-Ror2 signaling between osteoblast-lineage cells and osteoclast precursors enhances osteoclastogenesis. *Nat Med* 18: 405-12

Major MB, Camp ND, Berndt JD, Yi XH, Goldenberg SJ, et al. 2007. Wilms tumor suppressor WTX negatively regulates WNT/beta-catenin signaling. *Science* 316: 1043-46

Malinsky M, Challis RJ, Tyers AM, Schiffels S, Terai Y, et al. 2015. Genomic islands of speciation separate cichlid ecomorphs in an East African crater lake. *Science* 350: 1493-8

Manceau M, Domingues VS, Mallarino R, Hoekstra HE. 2011. The developmental role of Agouti in color pattern evolution. *Science* 331: 1062-5

Marques DA, Lucek K, Meier JI, Mwaiko S, Wagner CE, et al. 2016. Genomics of Rapid Incipient Speciation in Sympatric Threespine Stickleback. *PLoS genetics* 12: e1005887

Martin A, Papa R, Nadeau NJ, Hill RI, Counterman BA, et al. 2012. Diversification of complex butterfly wing patterns by repeated regulatory evolution of a Wnt ligand. *Proceedings of the National Academy of Sciences of the United States of America* 109: 12632-37

Martin SH, Dasmahapatra KK, Nadeau NJ, Salazar C, Walters JR, et al. 2013. Genome-wide evidence for speciation with gene flow in Heliconius butterflies. *Genome Res* 23: 1817-28

Matrone G, Wilson KS, Maqsood S, Mullins JJ, Tucker CS, Denvir MA. 2015. CDK9 and its repressor LARP7 modulate cardiomyocyte proliferation and response to injury in the zebrafish heart. *J Cell Sci* 128: 4560-71

Matsuo N, Tanaka S, Yoshioka H, Koch M, Gordon MK, Ramirez F. 2008. Collagen XXIV (Col24a1) gene expression is a specific marker of osteoblast differentiation and bone formation. *Connect Tissue Res* 49: 68-75

Mayr E. 1942. *Systematics and the origin of species*. Cambridge: Harvard University Press.

Mayr E. 1963. *Animal species and evolution*. Cambridge, MA: Belknap Press.

McGee MD, Schluter D, Wainwright PC. 2013. Functional basis of ecological divergence in sympatric stickleback. *Bmc Evol Biol* 13

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, et al. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 20: 1297-303

McKinnon JS, Rundle HD. 2002. Speciation in nature: the threespine stickleback model systems. *Trends in ecology & evolution* 17: 480-88

McPhail JD. 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: 1402-08

McPhail JD. 1992. Ecology and evolution of sympatric sticklebacks (Gasterosteus): Evidence for a species pair in Paxton Lake, Texada Island, British Columbia *Can J Zool* 70: 361-69

McPhail JD. 1993. Ecology and evolution of sympatric sticklebacks (Gasterosteus): Origins of the species pairs. *Can J Zool* 71: 515-23

McPhail JD. 1994. Speciation and the evolution of reproductive isolation in the sticklebacks (Gasterosteus) of south-western British Columbia In *The evolutionary biology of the threespine stickleback*, ed. MA Bell, SA Foster. Oxford: Oxford University Press

McVean G. 2007. The structure of linkage disequilibrium around a selective sweep. *Genetics* 175: 1395-406

Messer PW, Petrov DA. 2013. Population genomics of rapid adaptation by soft selective sweeps. *Trends in ecology & evolution* 28: 659-69

Meyer A, Kocher TD, Basasibwaki P, Wilson AC. 1990. Monophyletic Origin of Lake Victoria Cichlid Fishes Suggested by Mitochondrial-DNA Sequences. *Nature* 347: 550-53

Miller CT, Beleza S, Pollen AA, Schluter D, Kittles RA, et al. 2007. cis-Regulatory changes in Kit ligand expression and parallel evolution of pigmentation in sticklebacks and humans. *Cell* 131: 1179-89

Muller HJ. 1942. Isolating Mechanisms, Evolution, and Temperature. *Biological Symposia* 6: 71-125

Muller HJ, Pontecorvo G. 1942. Recessive genes causing interspecific sterility and other disharmonies between Drosophila melanogaster and simulans. *Genetics* 27: 157

Muto A, Orger MB, Wehman AM, Smear MC, Kay JN, et al. 2005. Forward genetic analysis of visual behavior in zebrafish. *PLoS genetics* 1: e66

Nachman MW, Hoekstra HE, D'Agostino SL. 2003. The genetic basis of adaptive melanism in pocket mice. *P Natl Acad Sci USA* 100: 5268-73

Nachman MW, Payseur BA. 2012. Recombination rate variation and speciation: theoretical predictions and empirical results from rabbits and mice.

Philosophical transactions of the Royal Society of London. Series B, Biological sciences 367: 409-21

Nadachowska-Brzyska K, Burri R, Olason PI, Kawakami T, Smeds L, Ellegren H. 2013. Demographic divergence history of pied flycatcher and collared flycatcher inferred from whole-genome re-sequencing data. *PLoS genetics* 9: e1003942

Nadeau NJ, Whibley A, Jones RT, Davey JW, Dasmahapatra KK, et al. 2012. Genomic islands of divergence in hybridizing Heliconius butterflies identified by large-scale targeted sequencing. *Philos T R Soc B* 367: 343-53

Nielsen R. 2005. Molecular signatures of natural selection. *Annual review of genetics* 39: 197-218

Nielsen R, Williamson S, Kim Y, Hubisz MJ, Clark AG, Bustamante C. 2005. Genomic scans for selective sweeps using SNP data. *Genome Res* 15: 1566-75

Noor MAF, Feder JL. 2006. Speciation genetics: evolving approaches. *Nature Reviews Genetics* 7: 851-61

Nordborg M, Charlesworth B, Charlesworth D. 1996. The effect of recombination on background selection. *Genetics research* 67: 159-74

Nosil P, Funk DJ, Ortiz-Barrientos D. 2009a. Divergent selection and heterogeneous genomic divergence. *Mol Ecol* 18: 375-402

Nosil P, Harmon LJ, Seehausen O. 2009b. Ecological explanations for (incomplete) speciation. *Trends in ecology & evolution* 24: 145-56

Nosil P, Vines TH, Funk DJ. 2005. Perspective: Reproductive isolation caused by natural selection against immigrants from divergent habitats. *Evolution; international journal of organic evolution* 59: 705-19

O'Brown NM, Summers BR, Jones FC, Brady SD, Kingsley DM. 2015. A recurrent regulatory change underlying altered expression and Wnt response of the stickleback armor plates gene EDA. *eLife* 4

Oksenberg N, Stevison L, Wall JD, Ahituv N. 2013. Function and Regulation of AUTS2, a Gene Implicated in Autism and Human Evolution. *PLoS genetics* 9

Orr HA. 2001. The genetics of species differences. *Trends in ecology & evolution* 16: 343-50

Orr HA. 2005. The genetic theory of adaptation: a brief history. *Nature reviews. Genetics* 6: 119-27

Pardo-Diaz C, Salazar C, Jiggins CD. 2015. Towards the identification of the loci of adaptive evolution. *Methods Ecol Evol* 6: 445-64

Park PJ. 2009. ChIP-seq: advantages and challenges of a maturing technology. *Nature reviews. Genetics* 10: 669-80

Pastinen T. 2010. Genome-wide allele-specific analysis: insights into regulatory variation. *Nature Reviews Genetics* 11: 533-38

Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, et al. 2012. Ancient admixture in human history. *Genetics* 192: 1065-93

Patterson N, Price AL, Reich D. 2006. Population structure and eigenanalysis. *PLoS genetics* 2: e190

Peichel CL, Nereng KS, Ohgi KA, Cole BLE, Colosimo PF, et al. 2001. The genetic architecture of divergence between threespine stickleback species. *Nature* 414: 901-05

Pennings PS, Hermisson J. 2006. Soft sweeps III: the signature of positive selection from recurrent mutation. *PLoS genetics* 2: e186

Peter BM, Huerta-Sanchez E, Nielsen R. 2012. Distinguishing between Selective Sweeps from Standing Variation and from a De Novo Mutation. *PLoS genetics* 8

Picelli S, Bjorklund AK, Reinius B, Sagasser S, Winberg G, Sandberg R. 2014. Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. *Genome Res* 24: 2033-40

Pickrell JK, Coop G, Novembre J, Kudaravalli S, Li JZ, et al. 2009. Signals of recent positive selection in a worldwide sample of human populations. *Genome Res* 19: 826-37

Pickrell JK, Pritchard JK. 2012. Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS genetics* 8: e1002967

Pickrell JK, Reich D. 2014. Toward a new history and geography of human genes informed by ancient DNA. *Trends in Genetics* 30: 377-89

Poelstra JW, Vijay N, Bossu CM, Lantz H, Ryll B, et al. 2014. The genomic landscape underlying phenotypic integrity in the face of gene flow in crows. *Science* 344: 1410-14

Pollard KS, Salama SR, Lambert N, Lambot MA, Coppens S, et al. 2006. An RNA gene expressed during cortical development evolved rapidly in humans. *Nature* 443: 167-72

Pool JE, Corbett-Detig RB, Sugino RP, Stevens KA, Cardeno CM, et al. 2012. Population Genomics of Sub-Saharan Drosophila melanogaster: African Diversity and Non-African Admixture. *PLoS genetics* 8

Popper AN, Ramcharitar J, Campana SE. 2005. Why otoliths? Insights from inner ear physiology and fisheries biology. *Mar Freshwater Res* 56: 497-504

Price MN, Dehal PS, Arkin AP. 2010. FastTree 2-Approximately Maximum-Likelihood Trees for Large Alignments. *Plos One* 5

Pritchard JK, Di Rienzo A. 2010. Adaptation - not by sweeps alone. *Nature reviews. Genetics* 11: 665-7

Protas ME, Hersey C, Kochanek D, Zhou Y, Wilkens H, et al. 2006. Genetic analysis of cavefish reveals molecular convergence in the evolution of albinism. *Nat Genet* 38: 107-11

Prud'homme B, Gompel N, Carroll SB. 2007. Emerging principles of regulatory evolution. *Proceedings of the National Academy of Sciences of the United States of America* 104 Suppl 1: 8605-12

Przeworski M, Coop G, Wall JD. 2005. The signature of positive selection on standing genetic variation. *Evolution; international journal of organic evolution* 59: 2312-23

Przeworski M, Hudson RR, Di rienzo A. 2000. Adjusting the focus on human variation. *Trends in Genetics* 16: 296-302

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, et al. 2007. PLINK: A tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics* 81: 559-75

Raeymaekers JAM, Konijnendijk N, Larmuseau MHD, Hellemans B, De Meester L, Volckaert FAM. 2014. A gene with major phenotypic effects as a target for selection vs. homogenizing gene flow. *Mol Ecol* 23: 162-81

Rahn JJ, Stackley KD, Chan SSL. 2013. Opa1 Is Required for Proper Mitochondrial Metabolism in Early Development. *Plos One* 8

Rai K, Jafri IF, Chidester S, James SR, Karpf AR, et al. 2010. Dnmt3 and G9a Cooperate for Tissue-specific Development in Zebrafish. *Journal of Biological Chemistry* 285: 4110-21

Rasmussen MD, Hubisz MJ, Gronau I, Siepel A. 2014. Genome-wide inference of ancestral recombination graphs. *PLoS genetics* 10: e1004342

Rebeiz M, Pool JE, Kassner VA, Aquadro CF, Carroll SB. 2009. Stepwise Modification of a Modular Enhancer Underlies Adaptation in a Drosophila Population. *Science* 326: 1663-67

Reed RD, Papa R, Martin A, Hines HM, Counterman BA, et al. 2011. optix Drives the Repeated Convergent Evolution of Butterfly Wing Pattern Mimicry. *Science* 333: 1137-41

Reid NM, Proestou DA, Clark BW, Warren WC, Colbourne JK, et al. 2016. The genomic landscape of rapid repeated evolutionary adaptation to toxic pollution in wild fish. *Science* 354: 1305-08

Reimchen TE. 1994. Predators and morphological evolution in threespine stickleback In *The evolutionary biology of the threespine stickleback*, ed. MA Bell, SA Foster, pp. 240-73. Oxford, U.K.: Oxford Univ. Press

Reimchen TE, Nosil P. 2006. Replicated ecological landscapes and the evolution of morphological diversity among Gasterosteus populations from an archipelago on the west coast of Canada. *Can J Zool* 84: 643-54

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: 2944-51

Renaut S, Grassa CJ, Yeaman S, Moyers BT, Lai Z, et al. 2013. Genomic islands of divergence are not affected by geography of speciation in sunflowers. *Nat Commun* 4

Reyes ML, Baker JA. 2017. The consequences of diet limitation in juvenile threespine stickleback: growth, lipid storage and the phenomenon of compensatory growth. *Ecol Freshw Fish* 26: 301-12

Ricard-Blum S. 2011. The Collagen Family. *Csh Perspect Biol* 3

Robu ME, Larson JD, Nasevicius A, Beiraghi S, Brenner C, et al. 2007. p53 activation by knockdown technologies. *PLoS genetics* 3: 787-801

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

Roesti M, Moser D, Berner D. 2013. Recombination in the threespine stickleback genome patterns and consequences. *Mol Ecol* 22: 3014-27

Roesti M, Salzburger W. 2014. Natural Selection: It's a Many-Small World After All. *Current biology : CB* 24: R959-62

Rudman SM, Schluter D. 2016. Ecological Impacts of Reverse Speciation in Threespine Stickleback. *Current biology : CB*

Rundle HD, Nagel L, Wenrick Boughman J, Schluter D. 2000. Natural selection and parallel speciation in sympatric sticklebacks. *Science* 287: 306-8

Rundle HD, Nosil P. 2005. Ecological speciation. *Ecol Lett* 8: 336-52

Rundle HD, Schluter D. 2004. Natural Selection and Ecological Speciation in Sticklebacks In *Adaptive Speciation*, ed. U Dieckmann, M Doebeli, JAJ Metz, D Tautz, pp. 192–209: Cambridge University Press

Rundle HD, Whitlock MC. 2001. A genetic interpretation of ecologically dependent isolation. *Evolution; international journal of organic evolution* 55: 198-201

Sabeti PC, Reich DE, Higgins JM, Levine HZ, Richter DJ, et al. 2002. Detecting recent positive selection in the human genome from haplotype structure. *Nature* 419: 832-7

Sabeti PC, Schaffner SF, Fry B, Lohmueller J, Varilly P, et al. 2006. Positive natural selection in the human lineage. *Science* 312: 1614-20

Savolainen O, Lascoux M, Merila J. 2013. Ecological genomics of local adaptation. *Nature Reviews Genetics* 14: 807-20

Schebesta M, Lien CL, Engel FB, Keating MT. 2006. Transcriptional profiling of caudal fin regeneration in zebrafish. *Thescientificworldjo* 6: 38-54

Schliewen UK, Tautz D, Paabo S. 1994. Sympatric Speciation Suggested by Monophyly of Crater Lake Cichlids. *Nature* 368: 629-32

Schluter D. 1993. Adaptive Radiation in Sticklebacks - Size, Shape, and Habitat Use Efficiency. *Ecology* 74: 699-709

Schluter D. 1994. Experimental-Evidence That Competition Promotes Divergence in Adaptive Radiation. *Science* 266: 798-801

Schluter D. 1995. Adaptive Radiation in Sticklebacks - Trade-Offs in Feeding Performance and Growth. *Ecology* 76: 82-90

Schluter D. 1998. Ecological causes of speciation In *Endless Forms: Species and Speciation*, ed. D Howard, S Berlocher. Oxford: Oxford University Press

Schluter D. 2000. Ecological character displacement in adaptive radiation. *American Naturalist* 156: S4-S16

Schluter D. 2001. Ecology and the origin of species. *Trends in ecology & evolution* 16: 372-80

Schluter D. 2009. Evidence for ecological speciation and its alternative. *Science* 323: 737-41

Schluter D, Conte GL. 2009. Genetics and ecological speciation. *Proceedings of the National Academy of Sciences of the United States of America* 106 Suppl 1: 9955-62

Schluter D, Marchinko KB, Barrett RDH, Rogers SM. 2010. Natural selection and the genetics of adaptation in threespine stickleback. *Philos T R Soc B* 365: 2479-86

Schluter D, McPhail JD. 1992. Ecological character displacement and speciation in sticklebacks. *The American naturalist* 140: 85-108

Schluter D, Nagel LM. 1995. Parallel Speciation by Natural-Selection. *American Naturalist* 146: 292-301

Schraiber JG, Akey JM. 2015. Methods and models for unravelling human evolutionary history. *Nature reviews. Genetics*

Schrider DR, Kern AD. 2016. S/HIC: Robust Identification of Soft and Hard Sweeps Using Machine Learning. *PLoS genetics* 12

Seehausen O. 2006. Conservation: losing biodiversity by reverse speciation. *Current biology : CB* 16: R334-7

Seehausen O, Butlin RK, Keller I, Wagner CE, Boughman JW, et al. 2014. Genomics and the origin of species. *Nature reviews. Genetics* 15: 176-92

Shapiro MD, Bell MA, Kingsley DM. 2006. Parallel genetic origins of pelvic reduction in vertebrates. *Proceedings of the National Academy of Sciences of the United States of America* 103: 13753-8

Sharan SK, Thomason LC, Kuznetsov SG, Court DL. 2009. Recombineering: a homologous recombination-based method of genetic engineering. *Nature Protocols* 4: 206-23

Sheykholeslami K, Kaga K. 2002. The otolithic organ as a receptor of vestibular hearing revealed by vestibular-evoked myogenic potentials in patients with inner ear anomalies. *Hearing Res* 165: 62-67

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

Sobel JM, Chen GF, Watt LR, Schemske DW. 2010. The biology of speciation. *Evolution; international journal of organic evolution* 64: 295-315

Sousa V, Hey J. 2013. Understanding the origin of species with genome-scale data: modelling gene flow. *Nature Reviews Genetics* 14: 404-14

Spence R, Wootton RJ, Barber I, Przybylski M, Smith C. 2013. Ecological causes of morphological evolution in the three-spined stickleback. *Ecology and evolution* 3: 1717-26

Spence R, Wootton RJ, Przybylski M, Zieba G, Macdonald K, Smith C. 2012. Calcium and salinity as selective factors in plate morph evolution of the three-spined stickleback (Gasterosteus aculeatus). *Journal of evolutionary biology* 25: 1965-74

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30: 1312-13

Stephan W. 2016. Signatures of positive selection: from selective sweeps at individual loci to subtle allele frequency changes in polygenic adaptation. *Mol Ecol* 25: 79-88

Stephan W, Charlesworth B, McVean G. 1999. The effect of background selection at a single locus on weakly selected, partially linked variants. *Genetical Research* 73: 133-46

Stern DL, Orgogozo V. 2008. The loci of evolution: How predictable is genetic evolution ? *Evolution; international journal of organic evolution* 62: 2155-77

Stern DL, Orgogozo V. 2009. Is Genetic Evolution Predictable? *Science* 323: 746-51

Tajima F. 1989. Statistical-Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism. *Genetics* 123: 585-95

Taylor EB, Boughman JW, Groenenboom M, Sniatynski M, Schluter D, Gow JL. 2006. Speciation in reverse: morphological and genetic evidence of the collapse of a three-spined stickleback (Gasterosteus aculeatus) species pair. *Mol Ecol* 15: 343-55

Taylor EB, McPhail JD. 1999. Evolutionary history of an adaptive radiation in species pairs of threespine sticklebacks (Gasterosteus): insights from mitochondrial DNA. *Biol J Linn Soc* 66: 271-91

Taylor EB, McPhail JD. 2000. Historical contingency and ecological determinism interact to prime speciation in sticklebacks, Gasterosteus. *Proceedings. Biological sciences / The Royal Society* 267: 2375-84

Terekhanova NV, Logacheva MD, Penin AA, Neretina TV, Barmintseva AE, et al. 2014. Fast Evolution from Precast Bricks: Genomics of Young Freshwater Populations of Threespine Stickleback Gasterosteus aculeatus. *PLoS genetics* 10: e1004696

Terhorst J, Kamm JA, Song YS. 2016. Robust and scalable inference of population history from hundreds of unphased whole genomes. *Nat Genet*

The Gene Ontology C. 2017. Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Res* 45: D331-D38

Tirosh I, Reikhav S, Levy AA, Barkai N. 2009. A Yeast Hybrid Provides Insight into the Evolution of Gene Expression Regulation. *Science* 324: 659-62

Tocher DR. 2003. Metabolism and functions of lipids and fatty acids in teleost fish. *Rev Fish Sci* 11: 107-84

Trapnell C, Roberts A, Goff L, Pertea G, Kim D, et al. 2012. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols* 7: 562-78

Tsao KC, Tu CF, Lee SJ, Yang RB. 2013. Zebrafish scube1 (Signal Peptide-CUB (Complement Protein C1r/C1s, Uegf, and Bmp1)-EGF (Epidermal Growth Factor) Domain-containing Protein 1) Is Involved in Primitive Hematopoiesis. *Journal of Biological Chemistry* 288: 5017-26

Tse WKF. 2017. Importance of deubiquitinases in zebrafish craniofacial development. *Biochem Bioph Res Co* 487: 813-19

Tse WKF, Eisenhaber B, Ho SHK, Ng Q, Eisenhaber F, Jiang YJ. 2009. Genome-wide loss-of-function analysis of deubiquitylating enzymes for zebrafish development. *BMC genomics* 10

Turner TL, Hahn MW, Nuzhdin SV. 2005. Genomic islands of speciation in Anopheles gambiae. *PLoS biology* 3: 1572-78

Urasaki A, Morvan G, Kawakami K. 2006. Functional dissection of the Tol2 transposable element identified the minimal cis-sequence and a highly repetitive sequence in the subterminal region essential for transposition. *Genetics* 174: 639-49

van Amerongen R, Fuerer C, Mizutani M, Nusse R. 2012. Wnt5a can both activate and repress Wnt/beta-catenin signaling during mouse embryonic development. *Dev Biol* 369: 101-14

van Doorn GS, Edelaar P, Weissing FJ. 2009. On the Origin of Species by Natural and Sexual Selection. *Science* 326: 1704-07

van't Hof AE, Edmonds N, Dalikova M, Marec F, Saccheri IJ. 2011. Industrial Melanism in British Peppered Moths Has a Singular and Recent Mutational Origin. *Science* 332: 958-60

Verta J, Jones FC. Adaptive transcriptomic divergence in sticklebacks has an additive cis-regulatory basis. *Submitted*

Via S. 2009. Natural selection in action during speciation. *Proceedings of the National Academy of Sciences of the United States of America* 106 Suppl 1: 9939-46

Via S. 2012. Divergence hitchhiking and the spread of genomic isolation during ecological speciation-with-gene-flow. *Philos T R Soc B* 367: 451-60

Vitti JJ, Grossman SR, Sabeti PC. 2013. Detecting natural selection in genomic data. *Annual review of genetics* 47: 97-120

Voight BF, Kudaravalli S, Wen X, Pritchard JK. 2006. A map of recent positive selection in the human genome. *PLoS biology* 4: e72

Wang M, Uebbing S, Ellegren H. 2017. Bayesian Inference of Allele-Specific Gene Expression Indicates Abundant Cis-Regulatory Variation in Natural Flycatcher Populations. *Genome Biology and Evolution* 9: 1266-79

Wang Y, Kaiser MS, Larson JD, Nasevicius A, Clark KJ, et al. 2010. Moesin1 and Ve-cadherin are required in endothelial cells during in vivo tubulogenesis. *Development* 137: 3119-28

Wark AR, Mills MG, Dang LH, Chan YF, Jones FC, et al. 2012. Genetic architecture of variation in the lateral line sensory system of threespine sticklebacks. *G3* 2: 1047-56

Wark AR, Peichel CL. 2010. Lateral line diversity among ecologically divergent threespine stickleback populations. *J Exp Biol* 213: 108-17

Webb JF, Smith WL, Ketten DR. 2006. *Fish Bioacoustics*. New York, NY: Springer.

Weissing FJ, Edelaar P, van Doorn GS. 2011. Adaptive speciation theory: a conceptual review. *Behav Ecol Sociobiol* 65: 461-80

Werner JD, Borevitz JO, Uhlenhaut NH, Ecker JR, Chory J, Weigel D. 2005a. FRIGIDA-independent variation in flowering time of natural Arabidopsis thaliana accessions. *Genetics* 170: 1197-207

Werner JD, Borevitz JO, Warthmann N, Trainer GT, Ecker JR, et al. 2005b. Quantitative trait locus mapping and DNA array hybridization identify an FLM deletion as a cause for natural flowering-time variation. *Proceedings of the National Academy of Sciences of the United States of America* 102: 2460-65

Westcot SE, Hatzold J, Urban MD, Richetti SK, Skuster KJ, et al. 2015. Protein-Trap Insertional Mutagenesis Uncovers New Genes Involved in Zebrafish Skin Development, Including a Neuregulin 2a-Based ErbB Signaling Pathway Required during Median Fin Fold Morphogenesis. *Plos One* 10

Westenberg M, Bamps S, Soedling H, Hope IA, Dolphin CT. 2010. Escherichia coli MW005: lambda Red-mediated recombineering and copy-number induction of oriV-equipped constructs in a single host. *Bmc Biotechnol* 10

White BJ, Cheng CD, Simard F, Costantini C, Besansky NJ. 2010. Genetic association of physically unlinked islands of genomic divergence in incipient species of Anopheles gambiae. *Mol Ecol* 19: 925-39

Wilson BA, Pennings PS, Petrov DA. 2017. Soft Selective Sweeps in Evolutionary Rescue. *Genetics* 205: 1573-86

Wittkopp PJ, Haerum BK, Clark AG. 2004. Evolutionary changes in cis and trans gene regulation. *Nature* 430: 85-8

Wittkopp PJ, Haerum BK, Clark AG. 2008. Regulatory changes underlying expression differences within and between Drosophila species. *Nat Genet* 40: 346-50

Wittkopp PJ, Kalay G. 2012. Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. *Nature Reviews Genetics* 13: 59-69

Wolf JBW, Ellegren H. 2017. Making sense of genomic islands of differentiation in light of speciation. *Nature Reviews Genetics* 18: 87-100

Wray GA. 2007a. The evolutionary significance of cis-regulatory mutations. *Nature Reviews Genetics* 8: 206-16

Wray GA. 2007b. The evolutionary significance of cis-regulatory mutations. *Nature reviews. Genetics* 8: 206-16

Xu D, Liu J, Fu T, Shan B, Qian L, et al. 2017. USP25 regulates Wnt signaling by controlling the stability of tankyrases. *Genes & development* 31: 1024-35

Xu F, Li K, Tian M, Hu P, Song W, et al. 2009. N-CoR is required for patterning the anterior-posterior axis of zebrafish hindbrain by actively repressing retinoid signaling. *Mech Develop* 126: 771-80

Yan H, Yuan WS, Velculescu VE, Vogelstein B, Kinzler KW. 2002. Allelic variation in human gene expression. *Science* 297: 1143-43

Yi X, Liang Y, Huerta-Sanchez E, Jin X, Cuo ZX, et al. 2010. Sequencing of 50 human exomes reveals adaptation to high altitude. *Science* 329: 75-8

Yong L, Thet Z, Zhu Y. 2017. Genetic editing of the androgen receptor contributes to impaired male courtship behavior in zebrafish. *J Exp Biol* 220: 3017-21

Zerbino DR, Achuthan P, Akanni W, Amode MR, Barrell D, et al. 2018. Ensembl 2018. *Nucleic Acids Res* 46: D754-D61

9 Appendix Information

Appendix Table 1. Sequencing coverage of benthics and limnetics from different lakes

Sample	Mean coverage	Sample	Mean coverage
PAXB04	12.05	**QRYB01**	14.24
PAXB05	13.54	**QRYB06**	9.43
PAXB06	11.37	**QRYB08**	9.93
PAXB07	13.34	**QRYB11**	11.37
PAXB08	12.89	**QRYB13**	13.5
PAXB09	11.57	**QRYB25**	12.69
PAXL01	17.63	**QRYL04**	10.13
PAXL05	10.95	**QRYL05**	12.26
PAXL09	12.95	**QRYL07**	11.65
PAXL10	10.5	**QRYL08**	12.95
PAXL13	11.85	**QRYL09**	10.93
PAXL14	14.02	**QRYL10**	12.8
PRIB02	9.82	**ENSB01**	11.79
PRIB05	10.6	**ENSB03**	9.3
PRIB06	9.3	**ENSB08**	14.01
PRIB07	18.96	**ENSB12**	13.68
PRIB11	10.03	**ENSB15**	13.3
PRIB15	14.91	**ENSB23**	13.8
PRIL16	17.36	**ENSL17**	28.15
PRIL17	12.26	**ENSL24**	14.2
PRIL18	9.11	**ENSL25**	13.48
PRIL102	22.07	**ENSL33**	13.44
PRIL108	25.25	**ENSL37**	12.79
PRIL112	22.59	**ENSL50**	12.1

Appendix Table 2. Sequencing coverage of additional Paxton Lake benthics and limnetics

Sample ID	Mean coverage	Sample ID	Mean Coverage
PAXB101	22.3	**PAXL126**	22.2
PAXB102	24.18	**PAXL128**	26.92
PAXB105	29.17	**PAXL129**	34.52
PAXB106	16.81	**PAXL130**	34.35
PAXB107	22.22	**PAXL131**	21.19
PAXB108	21.3	**PAXL132**	14.13
PAXB109	16.97	**PAXL133**	17.8
PAXB110	17.65	**PAXL138**	21.32
PAXB111	14.69	**PAXL139**	28.3
PAXB112	17.26	**PAXL140**	25.26
PAXB115	29.72	**PAXL141**	22.45
PAXB117	14.7	**PAXL144**	19.81
PAXB119	26.84	**PAXL145**	20.89
PAXB120	55.11	**PAXL147**	18.95
PAXB122	23.76	**PAXL148**	86.51
PAXB123	24.12	**PAXL149**	63.88
PAXB125	19.45	**PAXL150**	31.87

Appendix Table 3. Detail information and sequencing coverage of hybrid zone marine and freshwater stickleback individuals

No.	Individual ID	Drainage	Latitude	Longitude	Country	Sex	Collection Year	Collector	Coverage
Pacific Marine									
1	LITC_1_2015#4	Little Campbell River	49.015	-122.783	Canada	female	2015	Jukka-Pekka Verta	18.38
2	LITC_1_2015#5	Little Campbell River	49.015	-122.783	Canada	female	2015	Jukka-Pekka Verta	24.72
3	LITC_1_2015#6	Little Campbell River	49.015	-122.783	Canada	female	2015	Jukka-Pekka Verta	30.41
4	LITC_1_2015#7	Little Campbell River	49.015	-122.783	Canada	female	2015	Jukka-Pekka Verta	17.91
5	LITC_1_2015#8	Little Campbell River	49.015	-122.783	Canada	female	2015	Jukka-Pekka Verta	19.93
6	LITC_1_2015#9	Little Campbell River	49.015	-122.783	Canada	female	2015	Jukka-Pekka Verta	16.77
7	BIGR|1_32|2007#01	Big River	39.302	-123.786	USA	female	2007	Felicity Jones	4.94
8	BIGR|1_32|2007#02	Big River	39.302	-123.786	USA	male	2007	Felicity Jones	6.05
9	BIGR_1_32_2007#03	Big River	39.304	-123.78	USA	female	2007	Felicity Jones	5.23
10	BIGR|3_63|2007#08	Big River	39.302	-123.786	USA	female	2007	Felicity Jones	4.87
11	BIGR|3_63|2007#14	Big River	39.302	-123.786	USA	female	2007	Felicity Jones	5.02
12	BNMA|X|2006#01	Bonsall Creek	48.885	-123.673	Canada	female	2006	Tim Vines	5.04
13	BNMA|X|2006#02	Bonsall Creek	48.885	-123.673	Canada	female	2006	Tim Vines	5.53
14	BNMA|X|2006#03	Bonsall Creek	48.885	-123.673	Canada	female	2006	Tim Vines	4.28
15	BNMA|X|2006#05	Bonsall Creek	48.885	-123.673	Canada	female	2006	Tim Vines	4.5
16	BNMA|X|2006#07	Bonsall Creek	48.885	-123.673	Canada	female	2006	Tim Vines	4.45
Pacific Freshwater									
17	LITC_28_2015#12	Little Campbell River	49.011	-122.625	Canada	female	2015	Felicity Jones	14.05
18	LITC_28_2015#13	Little Campbell River	49.011	-122.625	Canada	female	2015	Felicity Jones	12.44
19	LITC_28_2015#14	Little Campbell River	49.011	-122.625	Canada	female	2015	Felicity Jones	9.54
20	LITC_28_2015#15	Little Campbell River	49.011	-122.625	Canada	female	2015	Felicity Jones	8.32
21	LITC_28_2015#16	Little Campbell River	49.011	-122.625	Canada	female	2015	Felicity Jones	16.26
22	LITC_28_2015#18	Little Campbell River	49.011	-122.625	Canada	female	2015	Felicity Jones	20.17

23	BIGR|52_54|2007#04	Big River	39.352	-123.558	USA	female	2007	Felicity Jones	4.81
24	BIGR|52_54|2007#05	Big River	39.352	-123.558	USA	female	2007	Felicity Jones	5.62
25	BIGR|52_54|2007#12	Big River	39.352	-123.558	USA	female	2007	Felicity Jones	5.44
26	BIGR|52_54|2007#17	Big River	39.352	-123.558	USA	female	2007	Felicity Jones	4.66
27	BIGR_52_54_2008#02	Big River	55.942	-2.788	USA	female	2008	Felicity Jones	5.58
28	BNST|X|2006#01	Bonsall Creek	48.876	-123.686	Canada	female	2006	Tim Vines	4.75
29	BNST|X|2006#06	Bonsall Creek	48.876	-123.686	Canada	female	2006	Tim Vines	4.91
30	BNST|X2006#08	Bonsall Creek	48.876	-123.686	Canada	female	2006	Tim Vines	4.91
31	BNST|X|2006#09	Bonsall Creek	48.876	-123.686	Canada	female	2006	Tim Vines	4.21
32	BNST|X|2006#10	Bonsall Creek	48.876	-123.686	Canada	male	2006	Tim Vines	5.08
Atlantic Marine									
33	TYNE_1_2001#02	River Tyne	56.009	-2.579	Scotland	female	2001	Felicity Jones	5.24
34	TYNE_1_2001#07	River Tyne	56.009	-2.579	Scotland	female	2001	Felicity Jones	8.66
35	TYNE_1_2001#08	River Tyne	56.009	-2.579	Scotland	female	2001	Felicity Jones	8.58
36	TYNE_1_2001#09	River Tyne	56.009	-2.579	Scotland	female	2001	Felicity Jones	8.5
37	TYNE_1_2001#10	River Tyne	56.009	-2.579	Scotland	female	2001	Felicity Jones	6.73
38	TYNE_1_2001#14	River Tyne	56.009	-2.579	Scotland	female	2001	Felicity Jones	8.93
39	MIDF|BDVW|2011#01	Midfjardara River	65.354	-20.912	Iceland	female	2011	Felicity Jones	6.29
40	MIDF|BDVW|2011#02	Midfjardara River	65.354	-20.912	Iceland	female	2011	Felicity Jones	5.11
41	MIDF|BLUP|2011#01	Midfjardara River	65.354	-20.912	Iceland	female	2011	Felicity Jones	5.65
42	MIDF|S101|2011#05	Midfjardara River	65.350	-20.911	Iceland	female	2011	Felicity Jones	6.54
43	MIDF|S101|2011#06	Midfjardara River	65.350	-20.911	Iceland	female	2011	Felicity Jones	5.37
Atlantic Freshwater									
44	TYNE_8_2003#902	River Tyne	55.942	-2.788	Scotland	female	2003	Felicity Jones	7.14
45	TYNE_8_2003#905	River Tyne	55.942	-2.788	Scotland	female	2003	Felicity Jones	5.24
46	TYNE_8_2003#906	River Tyne	55.942	-2.788	Scotland	female	2003	Felicity Jones	11.82
47	TYNE_8_2003#908	River Tyne	55.942	-2.788	Scotland	female	2003	Felicity Jones	9.07

48	TYNE_8_2003#919	River Tyne	55.942	-2.788	Scotland	female	2003	Felicity Jones	8.53
49	TYNE_8_2003#920	River Tyne	55.942	-2.788	Scotland	female	2003	Felicity Jones	8.58
50	MIDF|REND|2011#01	Midfjardara River	65.318	-20.897	Iceland	female	2011	Felicity Jones	6.68
51	MIDF|REND|2011#04	Midfjardara River	65.318	-20.897	Iceland	female	2011	Felicity Jones	4.98
52	MIDF|REND|2011#05	Midfjardara River	65.318	-20.897	Iceland	male	2011	Felicity Jones	5.76
53	MIDF|REND|2011#06	Midfjardara River	65.318	-20.897	Iceland	male	2011	Felicity Jones	6.23
54	MIDF|REND|2011#10	Midfjardara River	65.318	-20.897	Iceland	female	2011	Felicity Jones	5.9

Appendix Table 4. Sequencing coverage of global marine and freshwater stickleback ecotypes (excluding marine and freshwater stickleback ecotypes from Little Campbell River and River Tyne)

No.	Sample	Mean coverage	No.	Sample	Mean coverage	No	Sample	Mean coverage
1	SAMN02864913	5.38	**53**	SAMN02866133	6.21	**105**	SAMN02866135	4.88
2	SAMN02864935	5.89	**54**	SAMN02864879	7.41	**106**	SAMN02781076	4.58
3	SAMN02781060	8.68	**55**	SAMN02864920	5.1	**107**	SAMN02864863	5.46
4	SAMN02864934	5.23	**56**	SAMN02869623	4.52	**108**	SAMN02781679	9.23
5	SAMN02864921	4.92	**57**	SAMN02864854	5.46	**109**	SAMN02870195	6.96
6	SAMN02864940	6.45	**58**	SAMN02864909	6.16	**110**	SAMN02864932	5.64
7	SAMN02781065	6.95	**59**	SAMN02866139	6.86	**111**	SAMN02864915	5.45
8	SAMN02781677	8.04	**60**	SAMN02864907	6.69	**112**	SAMN02864918	6.08
9	SAMN02864906	5.48	**61**	SAMN02864901	7.06	**113**	SAMN02864930	6.33
10	SAMN02864894	4.86	**62**	SAMN02864896	5.69	**114**	SAMN02864908	5.15
11	SAMN02781675	9.2	**63**	SAMN02870194	6.09	**115**	SAMN02864914	5.72
12	SAMN02781684	7.85	**64**	SAMN02869634	6.42	**116**	SAMN02869644	4.55
13	SAMN02869630	5.24	**65**	SAMN02864888	5.51	**117**	SAMN02864938	6.26
14	SAMN02781114	8.52	**66**	SAMN02869622	6.32	**118**	SAMN02866141	6.45
15	SAMN02864936	6.27	**67**	SAMN02864941	5.39	**119**	SAMN02866146	4.9
16	SAMN02864862	5.32	**68**	SAMN02864927	5.65	**120**	SAMN02864873	5.07
17	SAMN02864857	4.61	**69**	SAMN02869647	5.6	**121**	SAMN02870196	6.61
18	SAMN02866140	4.53	**70**	SAMN02864865	5.64	**122**	SAMN02864926	5.51
19	SAMN02869626	4.86	**71**	SAMN02869627	5.12	**123**	SAMN02864911	6.31
20	SAMN02864876	5.35	**72**	SAMN02869631	5.32	**124**	SAMN02864856	4.7
21	SAMN02864849	5.46	**73**	SAMN02866147	6.34	**125**	SAMN02864895	5.68
22	SAMN02864877	4.94	**74**	SAMN02864922	6.64	**126**	SAMN02864858	6.15

23	SAMN02864923	4.49
24	SAMN02781108	7.86
25	SAMN02781101	9.32
26	SAMN02866142	5.31
27	SAMN02781686	10.76
28	SAMN02781075	9.62
29	SAMN02781682	10.09
30	SAMN02781077	9.29
31	SAMN02866137	6.46
32	SAMN02781062	9.47
33	SAMN02864929	5.71
34	SAMN02864919	5.76
35	SAMN02781089	9.27
36	SAMN02781688	8.09
37	SAMN02864853	5.65
38	SAMN02864910	5.57
39	SAMN02864867	5.03
40	SAMN02781097	11.43
41	SAMN02864939	4.67
42	SAMN02864864	5.64
43	SAMN02864866	5.02
44	SAMN02866132	5.71
45	SAMN02864878	4.94
46	SAMN02864861	5.24
47	SAMN02864904	5.72
48	SAMN02866136	4.78
49	SAMN02869632	5.48
75	SAMN02866138	5.5
76	SAMN02864892	5.08
77	SAMN02866144	5.73
78	SAMN02864925	5.1
79	SAMN02869640	6.12
80	SAMN02869641	4.73
81	SAMN02864870	5.39
82	SAMN02864893	5.67
83	SAMN02864902	6.08
84	SAMN02864874	5.45
85	SAMN02864850	6.21
86	SAMN02869629	5.16
87	SAMN02864937	4.82
88	SAMN02869642	7.08
89	SAMN02869635	8.1
90	SAMN02864916	6.7
91	SAMN02864846	5.71
92	SAMN02869645	6.33
93	SAMN02864917	7.34
94	SAMN02781070	9.94
95	SAMN02781695	7
96	SAMN02781110	6.84
97	SAMN02864855	5.03
98	SAMN02864890	5
99	SAMN02864900	5.69
100	SAMN02869633	6.05
101	SAMN02864844	5.24
127	SAMN02781079	9.77
128	SAMN02781080	5.53
129	SAMN02869624	5.48
130	SAMN02781073	7.92
131	SAMN02864848	5.91
132	SAMN02866131	5.33
133	SAMN02864875	5.82
134	SAMN02864897	5.23
135	SAMN02781084	8.88
136	SAMN02781102	9.12
137	SAMN02781693	9.11
138	SAMN02781681	9.3
139	SAMN02864871	4.97
140	SAMN02781096	5.29
141	SAMN02866143	6.23
142	SAMN02781100	8.31
143	SAMN02864912	5.44
144	SAMN02864869	4.93
145	SAMN02864931	5.01
146	SAMN02864928	6.09
147	SAMN02864860	5.9
148	SAMN02864889	5.44
149	SAMN02864933	7.03
150	SAMN02864872	5.99
151	SAMN02864942	5.54
152	SAMN02781098	5.6
153	SAMN02781071	4.43

50	SAMN02864891	7.6	**102**	SAMN02781103	6.17	**154**	SAMN02781696	5.32
51	SAMN02864852	5.24	**103**	SAMN02869636	4.37	**155**	SAMN02866145	5.34
52	SAMN02864924	5.03	**104**	SAMN02864859	4.77	**156**	SAMN02864905	4.84

Appendix Table 5. Number and size of parallel divergent regions between benthics and limnetics of each chromosome

Chromosome	Numbers of parallel divergent regions	Total size of parallel divergent regions	Chromosome size	Percentage of parallel divergent regions on the chromosome
chrI	413	9,456,087	28,185,914	33.55%
chrII	238	3,490,762	23,295,652	14.98%
chrIII	93	1,059,407	16,798,506	6.31%
chrIV	632	11,456,368	32,632,948	35.11%
chrV	66	1,123,934	12,251,397	9.17%
chrVI	68	646,932	17,083,675	3.79%
chrVII	439	11,663,061	27,937,443	41.75%
chrVIII	246	4,288,754	19,368,704	22.14%
chrIX	266	3,050,234	20,249,479	15.06%
chrX	95	1,106,905	15,657,440	7.07%
chrXI	204	2,541,796	16,706,052	15.21%
chrXII	187	3,216,313	18,401,067	17.48%
chrXIII	109	1,278,391	20,083,130	6.37%
chrXIV	17	59,483	15,246,461	0.39%
chrXV	8	45,492	16,198,764	0.28%
chrXVI	73	739,427	18,115,788	4.08%
chrXVII	75	1,506,425	14,603,141	10.32%
chrXVIII	91	2,029,409	16,282,716	12.46%
chrXIX	279	5,993,721	20,240,660	29.61%
chrXX	189	4,063,811	19,732,071	20.59%
chrXXI	166	1,999,834	11,717,487	17.07%
chrUn	371	6,263,129		

Appendix Table 6. "Islands of divergence" in the genome of benthics and limnetics

No	Chromosome	Start Position	End Position	Length
1	chrI	7,013,001	7,280,500	267,499
2	chrI	7,809,501	8,063,500	253,999
3	chrI	9,884,501	10,295,500	410,999
4	chrI	11,221,501	11,890,500	668,999
5	chrI	15,581,501	15,943,000	361,499
6	chrI	15,944,501	16,214,000	269,499
7	chrI	16,214,501	16,749,500	534,999
8	chrVII	6,551,001	6,814,500	263,499
9	chrVII	8,142,001	8,441,000	298,999
10	chrVII	9,642,001	10,173,000	530,999
11	chrVII	12,123,501	12,450,500	326,999
12	chrVII	14,531,501	15,216,500	684,999
13	chrVII	15,487,501	15,789,500	301,999
14	chrVII	17,281,001	17,739,000	457,999
15	chrVII	17,803,501	18,127,500	323,999
16	chrIX	11,593,001	11,942,500	349,499
17	chrXVII	7,632,001	7,925,500	293,499
18	chrXVIII	4,957,001	5,251,500	294,499
19	chrXIX	9,173,501	9,440,500	266,999
20	chrXIX	10,835,001	11,087,500	252,499
21	chrXIX	13,642,501	13,909,500	266,999
22	chrUn	1,507,501	1,828,500	320,999
23	chrUn	2,252,501	2,758,500	505,999
24	chrUn	3,492,001	3,783,000	290,999
25	chrUn	4,254,501	4,522,500	267,999

Appendix Table 7. “Islands of divergence” in the genome of Paxton Lake benthics and limnetics

No	Chromosome	Start	End	Length
1	chrI	6,170,001	6,986,500	816,499
2	chrI	7,416,001	8,393,000	976,999
3	chrI	8,719,501	9,371,000	651,499
4	chrI	9,620,001	10,334,500	714,499
5	chrI	10,732,501	11,968,500	1,235,999
6	chrI	15,608,001	16,202,500	594,499
7	chrI	16,214,501	16,920,500	705,999
8	chrIV	11,271,001	11,887,500	616,499
9	chrIV	24,893,001	25,455,000	561,999
10	chrIV	27,319,001	27,960,000	640,999
11	chrVII	6,172,501	6,810,000	637,499
12	chrVII	8,086,001	9,045,500	959,499
13	chrVII	9,576,001	10,179,000	602,999
14	chrVII	12,101,001	12,607,000	505,999
15	chrVII	13,778,001	14,337,000	558,999
16	chrVII	14,337,501	15,847,500	1,509,999
17	chrVII	16,296,001	16,876,000	579,999
18	chrVII	17,230,501	17,731,500	500,999
19	chrVII	18,366,501	18,913,500	546,999
20	chrVIII	7,636,501	8,599,500	962,999
21	chrXII	12,677,501	14,052,500	1,374,999
22	chrXIII	11,829,501	12,477,000	647,499
23	chrXVII	7,442,001	8,123,000	680,999
24	chrXIX	10,502,001	11,119,500	617,499
25	chrXX	9,005,001	9,628,500	623,499
26	chrXX	9,719,001	10,317,000	597,999
27	chrXXI	1,499,001	2,464,500	965,499
28	chrUn	760,001	1,797,500	1,037,499
29	chrUn	2,017,001	2,700,500	683,499
30	chrUn	3,369,001	4,123,500	754,499
31	chrUn	7,046,001	7,596,500	550,499
32	chrUn	8,401,001	9,080,500	679,499

Appendix Table 8. Genomic regions having extreme CSS and CLR scores in Paxton Lake Benthics

No.	Chromosome	Start	End	Length
1	chrIV	12,214,501	12,219,000	4,499
2	chrIV	15,737,001	15,739,500	2,499
3	chrIV	21,236,501	21,241,500	4,999
4	chrIV	21,265,501	21,269,500	3,999
5	chrIV	21,276,001	21,278,500	2,499
6	chrIV	21,290,501	21,294,500	3,999
7	chrIV	24,120,001	24,127,000	6,999
8	chrIV	24,155,001	24,159,000	3,999
9	chrIV	24,163,501	24,179,000	15,499
10	chrIV	24,191,001	24,200,000	8,999
11	chrIV	24,212,001	24,222,000	9,999
12	chrIV	24,415,001	24,418,000	2,999
13	chrIV	24,423,001	24,427,000	3,999
14	chrIV	24,429,501	24,433,000	3,499
15	chrIV	24,441,501	24,448,000	6,499
16	chrIV	24,466,501	24,472,000	5,499
17	chrIV	24,475,001	24,477,500	2,499
18	chrIV	24,486,501	24,490,000	3,499
19	chrIV	24,495,501	24,501,000	5,499
20	chrIV	24,502,001	24,514,500	12,499
21	chrIV	24,518,501	24,521,000	2,499
22	chrIV	24,537,501	24,544,500	6,999
23	chrIV	24,597,501	24,603,500	5,999
24	chrIV	24,807,501	24,812,000	4,499
25	chrIV	24,822,001	24,827,000	4,999
26	chrIV	24,935,001	24,948,500	13,499
27	chrIV	24,951,501	24,959,000	7,499
28	chrIV	24,960,001	24,971,500	11,499
29	chrIV	24,975,001	24,983,000	7,999
30	chrIV	24,982,501	24,986,000	3,499
31	chrIV	24,986,501	24,990,500	3,999
32	chrIV	24,991,501	24,997,000	5,499
33	chrIV	24,998,501	25,003,000	4,499
34	chrIV	25,003,501	25,018,000	14,499
35	chrIV	25,022,501	25,030,000	7,499
36	chrIV	25,030,501	25,053,000	22,499
37	chrIV	25,054,501	25,060,000	5,499
38	chrIV	25,060,501	25,069,000	8,499
39	chrIV	25,068,001	25,070,500	2,499
40	chrIV	25,073,001	25,079,000	5,999
41	chrIV	25,083,501	25,086,500	2,999
42	chrIV	25,203,001	25,205,500	2,499
43	chrIV	25,209,501	25,219,000	9,499
44	chrIV	25,263,501	25,291,000	27,499
45	chrIV	25,291,501	25,296,000	4,499
46	chrIV	25,299,501	25,326,500	26,999
47	chrIV	25,325,501	25,331,500	5,999
48	chrVII	14,036,501	14,041,000	4,499
49	chrVIII	9,233,001	9,235,500	2,499
50	chrVIII	9,245,001	9,248,000	2,999
51	chrVIII	9,255,001	9,258,000	2,999
52	chrVIII	9,260,001	9,262,500	2,499
53	chrVIII	9,267,001	9,270,500	3,499
54	chrVIII	9,275,001	9,279,500	4,499
55	chrVIII	9,299,001	9,303,500	4,499
56	chrUn	4,191,001	4,197,000	5,999
57	chrUn	4,254,001	4,258,000	3,999

Appendix Table 9. Genomic regions having extreme CSS and CLR scores in Paxton Lake Limnetics

No	Chromosome	Start	End	Length	No	Chromosome	Start	End	Length
1	chrIV	19,963,501	19,967,000	3,499	**24**	chrVIII	7,057,501	7,060,500	2,999
2	chrIV	20,169,501	20,172,000	2,499	**25**	chrVIII	7,074,501	7,077,500	2,999
3	chrIV	20,189,501	20,195,500	5,999	**26**	chrVIII	7,090,501	7,094,500	3,999
4	chrIV	20,198,501	20,206,000	7,499	**27**	chrVIII	7,101,501	7,109,500	7,999
5	chrIV	20,375,001	20,379,500	4,499	**28**	chrVIII	8,110,001	8,112,500	2,499
6	chrIV	20,922,501	20,932,000	9,499	**29**	chrVIII	8,114,001	8,144,500	30,499
7	chrIV	20,933,501	20,939,500	5,999	**30**	chrVIII	8,144,001	8,146,500	2,499
8	chrIV	23,881,501	23,888,500	6,999	**31**	chrVIII	8,148,501	8,152,500	3,999
9	chrIV	23,928,001	23,932,000	3,999	**32**	chrVIII	8,169,001	8,172,000	2,999
10	chrIV	23,944,001	23,950,000	5,999	**33**	chrVIII	8,182,501	8,191,000	8,499
11	chrIV	23,953,501	23,959,000	5,499	**34**	chrVIII	8,196,501	8,200,000	3,499
12	chrIV	23,993,001	23,996,000	2,999	**35**	chrVIII	8,320,501	8,324,000	3,499
13	chrIV	24,031,001	24,039,000	7,999	**36**	chrVIII	8,330,501	8,335,000	4,499
14	chrIV	24,116,501	24,127,000	10,499	**37**	chrVIII	8,355,001	8,358,500	3,499
15	chrIV	24,133,501	24,144,000	10,499	**38**	chrVIII	8,361,501	8,368,000	6,499
16	chrIV	24,146,501	24,151,000	4,499	**39**	chrVIII	8,369,501	8,377,500	7,999
17	chrIV	24,151,501	24,157,500	5,999	**40**	chrVIII	8,381,501	8,386,000	4,499
18	chrIX	9,298,501	9,302,500	3,999	**41**	chrVIII	8,573,001	8,575,500	2,499
19	chrVII	17,145,501	17,148,000	2,499	**42**	chrUn	1,435,001	1,438,000	2,999
20	chrVII	17,149,501	17,153,000	3,499	**43**	chrUn	1,481,501	1,486,000	4,499
21	chrVII	17,183,001	17,186,000	2,999	**44**	chrUn	1,504,001	1,506,500	2,499
22	chrVII	17,191,501	17,196,000	4,499	**45**	chrUn	1,508,501	1,515,000	6,499
23	chrVII	17,199,001	17,203,000	3,999					

Appendix Table 10. Primers used for SNP validation

No.	Target SNPs	Left Primers	Right Primers
1	chrI:1825326	GATAAACGTCCCACTGTGCC	CCTGAAGGGTCGCATAATAGG
2	chrI:8940697	ACAGGGCAGTGAGAGACAGG	ATGTAAAGATGGCACCTCGG
3	chrI:9025008	TTGCTCCAGACATATCAGTCG	CTCCATCACTCCAACAATCC
4	chrI:10109245	GCTCTGCATTGACAGGACG	TGGTTAAGGATAACGTCGCC
5	chrI:11950371	TTGATTCCCACCTTTGATCC	CATCTGGGTCGACATTTGC
6	chrI:20434005	CAAAGCAGATAACACGTGGC	AACACTGGCTGACATGAAAGG
7	chrII:1863102	GCATGGATATGCCACAAGC	AGGACACTCAGAGCACAAGC
8	chrII:6208664	CATCGAGTCTGTGAGCAGCC	TTTAAAGCGGTGTGACGC
9	chrII:7157131	CGCCTTCACTCATTCTGTCC	CAGCAGACTGTGGTAATATCTCG
10	chrII:20750645	ACACGCGTCAAGGGTGTATT	CTTCGACCATATCGCCTCAT
11	chrIII:1528701	AGCAGCATTGTTCATAACGG	GAGGGCAGTGACAGCAGC
12	chrIII:2406134	GATGTCTGCAAAGGTGATGG	CGAGTCTGCACTCATGAACC
13	chrIII:13150896	CAGTTCATAAGCGGTTCTTCC	TGTTTGGGTGACCGGAGG
14	chrIV:154648	TGAACAATGTCTCTCTGAACGG	CCGAGGTACTCTCCTCCTCC
15	chrIV:213239	TCATAAGCTCAGACCCTCCG	ACATCACAGGAAGTGACGCC
16	chrIV:1462987	TCCCATCTAATGCTGTAACGC	GAAGTTACGCCTCATGGACC
17	chrIV:1962133	CCTCGTGTTAATGCATCGG	TCTCCTGTGAGGACGAATCC
18	chrIV:2832425	TAGATGGCAGAACAACACCG	GGTCCTTGTGATTGATGCG
19	chrIV:30446328	TGTTGTTGTTCAGAGGTGGC	CTTGGTCTTGATGCCTTTCC
20	chrV:6850659	TCAGACCCACGAGTTATCCC	GGAAGTATGCAGAGGAAGGG
21	chrVI:14686732	TAAGCATTGATCTTGTGCCC	GAAGCAGGTTAAGAGGCAGG
22	chrVI:15032557	AGACAGAGGAGCCCATCAGC	GCAACATAATGGGACAAGCC
23	chrVI:15980632	GGTGAAGACACAAAGGGTGG	AATTGTGAGTCATTCGTGCG
24	chrVI:16771234	CCACTGTCTTTATCCGCACC	TGAGGTCTGTGGATGACACC
25	chrVII:3258420	TGTTAGATCCACCTGCCTGG	TAACCTGTTCCGTCTCCTGC
26	chrVII:4410387	TGAGTTACACATAAGACAGGCCC	ATTAAGCGTGCATGAGTTCC
27	chrVII:7138847	CACATTGTAATGGAGATGCCC	CTGGAGAAGGAACGTCAAGC
28	chrVII:8253857	AGTAGTCATGAAACTGCTGCG	CAGAATGTGTAACTGTTCCTGC
29	chrVII:15568229	AAACGTCCATGTTTGCTGC	TTCATACAGAGATGCTGCCG
30	chrVII:23854348	TCTATCACGTGACGCTGTGG	TCGTTAGTGAGACAGCTGGG
31	chrVII:25086503	GGACATGTGATACAGCCCG	CTCTGAGCGTTGTTCCTGC
32	chrVIII:5683221	CCATGTCGAGTAAGTGTGGC	TCTTTGCTGAAACCCTTTCC
33	chrVIII:7076647	GCTGTATTACATCGACGTGGC	TTAACAAACGGGTTGATGGG
34	chrVIII:19110568	AATCTGTCAGAGGGACGAGC	TTCTGGAACCACACACCTCC
35	chrIX:6226794	AACAGCATGCAGACAAGTGC	CTGATAGATGCGTGATAGCTGC
36	chrIX:10405007	CGTTTGGATCTTTCCTCTGC	TGTTTAAATGGTACCGTGGC
37	chrIX:10874438	GTTAAGGCTACCATCCTGGC	GCGCACACACACTTACGC
38	chrIX:10958964	GGAGTGAACTGCATGATTCG	TCAGTTCTACGCCAGCACC
39	chrIX:17932100	GAATAATCTGTGCGCAGTGG	GAAATATTCCCTCCGCTGG
40	chrIX:19215561	GTGTTAATACCGTCCACCCG	GTGAGCGACCTCACTGTACC
41	chrIX:19815079	ACAACATCCTGTAGAGGCCC	TGTGCTCATTGGTTGAGTGC
42	chrX:6796918	AGGTCTGCAAACTACGTCGC	GCTAGCTGGTTAGCCGAGG
43	chrXI:2862121	GATCACAACATGCTCCCTCC	TTACAGTGTGATGGACGACG
44	chrXI:3180424	AACAGAGGGAAGGGAGAAGC	CATTCGATTGAGTGAAGCCC
45	chrXI:3974245	AAGGCATTGTGTGAAGGAGG	TGCTCAGAACTCATTGCTCG
46	chrXI:7012389	CTTTCTCAACGCTCTCCAGC	CGGACTGTACGAGTGAGAAGC

47	chrXI:8637644	GTGGCTGATGTTAATGCAGG	CACGGCTGTGTTAGAGAAGC
48	chrXI:9711230	CGTTCCCTGAAGTGAAAGC	CATTCAAATGCTTCACAGGC
49	chrXI:13518171	GCAGTTTCGTTTGTGAATGG	TGCCCTTGTATTTGTCAACG
50	chrXII:1997378	ATTGAAGCAGCAACAAAGGC	TATGCAGCAGCATTAGAGGG
51	chrXII:2182641	AACAAAGTGTGCCCTATGCC	GAGACCAGATGAAGGCTCCC
52	chrXII:4813342	CCCTGTATATGTTGGTGTCCC	GCAATTTGTGGAATGTGCG
53	chrXII:4915843	AGCCTGCTAGCGTCATAACC	GTCAACTGAGGTTGCAGTCG
54	chrXII:6986471	GAGAGGGAGGCTACACCTGC	TGTGTTACAGGTAGAGAGACGGG
55	chrXII:11892442	CAATGCAGATCCAGGTGC	GGCCACACAGTGGAGTGC
56	chrXII:12442931	GTGACACATTTGAGGCTTGC	GTCCTCTAAATGCCTCGTGC
57	chrXII:12808143	CAAACAGCCAGAAGAATGGC	TAAGGAAATCATTGGGAGGC
58	chrXIII:590641	AGGTAGTGAGTGGGTGGTGG	GATTTCCTGGAGAGAGAACCC
59	chrXIII:6209100	ATTCCAAGACGATAATGCCG	AAAGTCTCACTGGAGCTGGG
60	chrXIV:2077894	ACTCCGCAGAGAGCAGAGG	CAACACACTGTTCCTTTCGC
61	chrXIV:2376483	CTGTTCATGAAGGTCAACGG	CACACTCTGCATCAAGTGGC
62	chrXIV:3883715	GCACTATTTCCTGCTTGTTGC	CACCATCGAAAGCAGTTTCC
63	chrXIV:4546291	AGAGACATTCCACCTCCACG	TATCTCCGTCTTGCGTGAGC
64	chrXIV:4563431	GGAGGGAAATTTGAATCCG	AATATTGGTCTCGTCGGTGC
65	chrXV:713347	GCTTCAGGTGGTCTTTGACG	GAAATTTCTCGCAGGCCC
66	chrXV:8725114	AAACAGCACACACATAAGCG	ACACTGCCTTACCTCCAACG
67	chrXV:9551006	GGAAGCAGATATAAACGGACG	CCCAATTCGCCACTATAAGC
68	chrXV:11165835	TCCCAGTAATCACGGAATGC	TTTATTGAGGAAACCCTGCG
69	chrXVI:15419555	CAGAGGAGTTCTCACCAGCC	GGCTAACGGTGCTAACGAGG
70	chrXVI:4279583	GATGCTGGTGACTTTCATGG	ACTTCCTGGGTTGATGTTCG
71	chrXVI:16163452	TTTCTTCCCTCTACATGCTGC	AGCCAGCGAGTTATGAGAGC
72	chrXVII:125860	AAAGGAGGAGATGCTGATGG	GAAGAAATGATGGTGCCTGG
73	chrXVII:2202841	TGTGGAGACCGACAATTTCC	GAGCTTAAATCATGACGCAGG
74	chrXVII:11517599	AACACACACGCATGCACC	TCCAGTTCATGCCGTTCC
75	chrXVIII:940689	TAGTGTTTGGATGTCGCACC	CCCTAACACACACCACTCCC
76	chrXVIII:11146361	CTCCACAAGACAGATGTGGG	AGGAGACAGAGACGGACTCG
77	chrXIX:124156	GTACCAGTGAAGAGAGCGGC	TTCCATGACCGTATGAACCC
78	chrXIX:5825824	CGATGGATCACACTGGAGG	CTGTTGTGTCGTCGTGAAGG
79	chrXIX:10978435	AGGACGTGAGAGAGTCGTGC	GTAAACAAACAGAAGGGCCG
80	chrXXI:4378964	CAAGTTGAGCAAATGCTTCG	CCAATTCCACAGTAATGGGC
81	chrXXI:9985787	GAGGTCTGCTTTGAGGACCC	GTTTCAAGAAGACAAGCGGC
82	chrXX:16213669	ACATTTCGATCGCTTCGC	TTTGAGCTCAGCATGTGTCC
83	chrXX:17577903	CAGGAATGTTCCCACAATGC	CGTCAGTGTCAGAAACCCG
84	chrUn:5463348	AATGAAGTAACGCCAGACGG	CTGAGGCTCATTCGAAATCC
85	chrUn:16395850	TGTCTAATCTTTGCGGCTCC	CTCCGTCCAATATCACTCACC
86	chrUn:16973303	TAAAGGTTTCCAGTGGTCCG	GAGATGCTGAACTCCAACCC
87	chrUn:17098174	TGTGCCTCTCCACCTACAGC	AACAACACGAGGGTACTCGG
88	chrUn:23813224	CCTCCAAGAGTCACACATGC	TGCTTTCAAAGACGTCATGC
89	chrUn:26458697	GCATTCAAGGATCATCAGGG	ACCTCAAACAGGGTCAGTCG
90	chrUn:26565105	TGTTTGTACTGATCCCATCTGC	AACACGACACGGACCTGG
91	chrUn:27739067	GTTTGTGTGTGTTTGTGTGTGC	TACTGCAGAGCTCCGATGG
92	chrUn:29412111	ACACGTTACTCCGTCATGGG	GAGCTTGTGACGTTAGCTGC
93	chrUn:40698981	CAGCTGACCACACAAACAGC	GGAGGCCAGGTAGAACTCG
94	chrUn:46131038	AAAGTTGGGTGAGACCAGG	GCCGCGTTGTATTGTAAAGG

Appendix Table 11. Adaptive regions of Paxton Lake benthics and limnetics

No	Chromosome	Start	End	Length	No.	Chromosome	Start	End	Length
1	chrI	7,573,501	7,577,500	3,999	**66**	chrVII	14,811,501	14,814,000	2,499
2	chrI	8,293,001	8,297,000	3,999	**67**	chrVII	14,814,501	14,817,500	2,999
3	chrI	9,464,501	9,467,500	2,999	**68**	chrVII	14,821,501	14,825,000	3,499
4	chrIV	14,406,001	14,408,500	2,499	**69**	chrVII	14,831,501	14,836,000	4,499
5	chrIV	15,255,001	15,258,500	3,499	**70**	chrVII	14,841,501	14,848,000	6,499
6	chrIV	19,206,501	19,209,000	2,499	**71**	chrVII	14,903,501	14,907,500	3,999
7	chrIV	19,212,501	19,216,500	3,999	**72**	chrVII	15,010,001	15,012,500	2,499
8	chrIV	20,189,501	20,193,500	3,999	**73**	chrVII	15,013,501	15,016,000	2,499
9	chrIV	20,517,001	20,523,500	6,499	**74**	chrVII	15,018,501	15,022,500	3,999
10	chrIV	24,120,001	24,127,000	6,999	**75**	chrVII	15,030,501	15,033,500	2,999
11	chrIV	24,155,001	24,157,500	2,499	**76**	chrVII	15,039,001	15,041,500	2,499
12	chrIV	25,302,501	25,307,000	4,499	**77**	chrVII	15,045,001	15,047,500	2,499
13	chrIV	25,318,001	25,321,500	3,499	**78**	chrVII	17,145,501	17,148,000	2,499
14	chrIV	25,337,501	25,341,500	3,999	**79**	chrVII	17,149,501	17,153,000	3,499
15	chrIV	27,977,001	27,979,500	2,499	**80**	chrVII	17,183,001	17,186,000	2,999
16	chrIX	8,531,001	8,535,000	3,999	**81**	chrVII	17,279,001	17,285,500	6,499
17	chrIX	9,819,501	9,826,000	6,499	**82**	chrVIII	7,053,501	7,057,500	3,999
18	chrIX	13,936,501	13,940,000	3,499	**83**	chrVIII	7,082,501	7,086,000	3,499
19	chrUn	1,481,501	1,486,000	4,499	**84**	chrVIII	7,090,501	7,094,500	3,999
20	chrUn	1,509,501	1,514,000	4,499	**85**	chrVIII	7,101,501	7,106,500	4,999
21	chrUn	1,640,001	1,642,500	2,499	**86**	chrVIII	7,109,501	7,113,500	3,999
22	chrUn	2,378,001	2,380,500	2,499	**87**	chrVIII	7,119,501	7,122,000	2,499
23	chrUn	5,188,501	5,192,000	3,499	**88**	chrVIII	7,260,001	7,265,000	4,999
24	chrUn	5,204,501	5,208,500	3,999	**89**	chrVIII	7,946,501	7,951,000	4,499
25	chrUn	5,216,501	5,220,000	3,499	**90**	chrVIII	8,114,001	8,119,000	4,999
26	chrUn	5,243,001	5,248,000	4,999	**91**	chrVIII	8,121,501	8,130,500	8,999
27	chrUn	5,610,001	5,615,000	4,999	**92**	chrVIII	8,187,001	8,190,500	3,499
28	chrUn	8,087,501	8,090,000	2,499	**93**	chrVIII	8,208,501	8,212,000	3,499
29	chrVII	6,739,001	6,742,000	2,999	**94**	chrVIII	8,330,501	8,335,000	4,499
30	chrVII	8,221,501	8,226,000	4,499	**95**	chrVIII	8,369,501	8,374,500	4,999
31	chrVII	8,230,001	8,232,500	2,499	**96**	chrVIII	8,381,501	8,386,000	4,499
32	chrVII	9,440,501	9,445,000	4,499	**97**	chrVIII	11,385,001	11,388,500	3,499
33	chrVII	9,867,001	9,871,500	4,499	**98**	chrVIII	11,642,001	11,646,000	3,999
34	chrVII	9,978,001	9,981,000	2,999	**99**	chrVIII	11,655,001	11,659,500	4,499
35	chrVII	10,006,001	10,010,000	3,999	**100**	chrVIII	11,711,001	11,715,000	3,999
36	chrVII	10,010,501	10,015,500	4,999	**101**	chrXI	9,041,001	9,046,500	5,499
37	chrVII	10,027,501	10,030,000	2,499	**102**	chrXI	9,061,501	9,067,000	5,499
38	chrVII	10,373,501	10,377,500	3,999	**103**	chrXII	11,119,001	11,123,000	3,999
39	chrVII	12,215,001	12,217,500	2,499	**104**	chrXII	12,125,501	12,129,500	3,999
40	chrVII	12,219,501	12,225,000	5,499	**105**	chrXII	12,237,501	12,240,500	2,999
41	chrVII	12,227,001	12,234,000	6,999	**106**	chrXII	12,268,501	12,271,000	2,499
42	chrVII	12,315,501	12,320,000	4,499	**107**	chrXII	12,311,501	12,315,000	3,499

43	chrVII	12,321,501	12,325,500	3,999	**108**	chrXII	12,327,501	12,330,500	2,999
44	chrVII	12,488,501	12,492,500	3,999	**109**	chrXII	12,552,001	12,554,500	2,499
45	chrVII	12,508,501	12,513,500	4,999	**110**	chrXII	12,574,501	12,578,000	3,499
46	chrVII	12,526,501	12,529,500	2,999	**111**	chrXII	12,601,501	12,605,000	3,499
47	chrVII	12,548,501	12,551,500	2,999	**112**	chrXII	12,608,501	12,611,000	2,499
48	chrVII	12,552,501	12,555,500	2,999	**113**	chrXII	12,616,001	12,619,000	2,999
49	chrVII	13,846,001	13,851,500	5,499	**114**	chrXII	12,800,501	12,806,500	5,999
50	chrVII	13,862,001	13,866,500	4,499	**115**	chrXII	12,898,001	12,901,000	2,999
51	chrVII	14,035,001	14,041,000	5,999	**116**	chrXII	13,210,001	13,213,000	2,999
52	chrVII	14,108,001	14,113,500	5,499	**117**	chrXII	13,216,501	13,219,000	2,499
53	chrVII	14,118,001	14,122,000	3,999	**118**	chrXII	13,282,501	13,287,000	4,499
54	chrVII	14,133,501	14,145,500	11,999	**119**	chrXII	13,299,501	13,302,500	2,999
55	chrVII	14,147,501	14,153,500	5,999	**120**	chrXII	13,311,001	13,314,000	2,999
56	chrVII	14,154,501	14,157,500	2,999	**121**	chrXII	13,577,501	13,581,000	3,499
57	chrVII	14,160,501	14,163,000	2,499	**122**	chrXII	13,586,001	13,590,500	4,499
58	chrVII	14,237,001	14,239,500	2,499	**123**	chrXIII	12,172,001	12,179,500	7,499
59	chrVII	14,655,001	14,657,500	2,499	**124**	chrXIX	10,525,001	10,529,000	3,999
60	chrVII	14,668,001	14,670,500	2,499	**125**	chrXIX	10,543,501	10,546,500	2,999
61	chrVII	14,746,001	14,751,500	5,499	**126**	chrXIX	12,655,501	12,658,500	2,999
62	chrVII	14,755,501	14,759,000	3,499	**127**	chrXIX	16,058,501	16,061,500	2,999
63	chrVII	14,760,001	14,763,000	2,999	**128**	chrXVI	5,983,501	5,987,500	3,999
64	chrVII	14,793,501	14,798,000	4,499	**129**	chrXVII	7,900,001	7,903,000	2,999
65	chrVII	14,798,001	14,806,000	7,999	**130**	chrXX	14,338,001	14,341,000	2,999

Appendix Table 12. “Strongly adaptive regions” of benthics and limnetics

No.	Chromsome	Start	End	Length	No.	Chromsome	Start	End	Length
1	chrI	8,293,501	8,297,000	3,499	**41**	chrVII	15,013,501	15,016,000	2,499
2	chrIV	20,189,501	20,193,500	3,999	**42**	chrVII	15,018,501	15,022,500	3,999
3	chrIV	20,518,501	20,523,000	4,499	**43**	chrVII	15,030,501	15,033,500	2,999
4	chrIV	25,319,001	25,321,500	2,499	**44**	chrVII	15,039,001	15,041,500	2,499
5	chrIX	9,819,501	9,826,000	6,499	**45**	chrVII	15,045,001	15,047,500	2,499
6	chrIX	13,936,501	13,940,000	3,499	**46**	chrVII	17,145,501	17,148,000	2,499
7	chrUn	1,509,501	1,513,500	3,999	**47**	chrVII	17,149,501	17,152,500	2,999
8	chrUn	2,378,001	2,380,500	2,499	**48**	chrVII	17,183,001	17,186,000	2,999
9	chrUn	8,087,501	8,090,000	2,499	**49**	chrVIII	7,110,501	7,113,000	2,499
10	chrVII	6,739,001	6,742,000	2,999	**50**	chrVIII	7,261,001	7,263,500	2,499
11	chrVII	8,221,501	8,226,000	4,499	**51**	chrVIII	8,115,501	8,119,000	3,499
12	chrVII	8,230,001	8,232,500	2,499	**52**	chrVIII	8,122,501	8,130,500	7,999
13	chrVII	9,440,501	9,445,000	4,499	**53**	chrVIII	8,187,501	8,190,500	2,999
14	chrVII	9,867,001	9,871,500	4,499	**54**	chrVIII	8,209,501	8,212,000	2,499
15	chrVII	9,978,001	9,981,000	2,999	**55**	chrVIII	8,331,001	8,334,500	3,499
16	chrVII	10,006,001	10,010,000	3,999	**56**	chrVIII	11,642,001	11,646,000	3,999
17	chrVII	10,010,501	10,015,500	4,999	**57**	chrVIII	11,655,001	11,659,500	4,499
18	chrVII	10,027,501	10,030,000	2,499	**58**	chrXI	9,061,501	9,067,000	5,499
19	chrVII	10,373,501	10,377,500	3,999	**59**	chrXII	11,119,001	11,123,000	3,999
20	chrVII	12,215,001	12,217,500	2,499	**60**	chrXII	12,125,501	12,129,500	3,999
21	chrVII	12,219,501	12,225,000	5,499	**61**	chrXII	12,552,001	12,554,500	2,499
22	chrVII	12,227,001	12,234,000	6,999	**62**	chrXII	12,574,501	12,578,000	3,499
23	chrVII	12,315,501	12,320,000	4,499	**63**	chrXII	12,601,501	12,605,000	3,499
24	chrVII	12,321,501	12,325,500	3,999	**64**	chrXII	12,608,501	12,611,000	2,499
25	chrVII	12,510,001	12,512,500	2,499	**65**	chrXII	12,616,001	12,619,000	2,999
26	chrVII	13,846,001	13,851,500	5,499	**66**	chrXII	12,800,501	12,806,500	5,999
27	chrVII	13,862,001	13,866,500	4,499	**67**	chrXII	12,898,501	12,901,000	2,499
28	chrVII	14,035,001	14,041,000	5,999	**68**	chrXII	13,216,501	13,219,000	2,499
29	chrVII	14,655,001	14,657,500	2,499	**69**	chrXII	13,282,501	13,287,000	4,499
30	chrVII	14,668,001	14,670,500	2,499	**70**	chrXII	13,299,501	13,302,500	2,999
31	chrVII	14,747,001	14,749,500	2,499	**71**	chrXII	13,311,001	13,314,000	2,999
32	chrVII	14,756,001	14,758,500	2,499	**72**	chrXII	13,577,501	13,581,000	3,499
33	chrVII	14,760,001	14,762,500	2,499	**73**	chrXII	13,586,001	13,590,500	4,499
34	chrVII	14,793,501	14,798,000	4,499	**74**	chrXIX	10,525,001	10,529,000	3,999
35	chrVII	14,798,001	14,806,000	7,999	**75**	chrXIX	10,543,501	10,546,500	2,999
36	chrVII	14,811,501	14,814,000	2,499	**76**	chrXIX	12,655,501	12,658,500	2,999
37	chrVII	14,821,501	14,825,000	3499	**77**	chrXVII	7,900,001	7,902,500	2,499
38	chrVII	14,831,501	14,836,000	4,499					
39	chrVII	14,841,501	14,848,000	6,499					
40	chrVII	15,010,001	15,012,500	2,499					

Appendix Table 13. Genetic divergence (F_{ST}) of benthic-marine, limnetic-marine, and benthic-limnetic ecotype pairs at adaptive loci of marine and freshwater sticklebacks across Northern Hemisphere

Chrom	Start	End	Benthic-Marine divergence (F_{ST})	Limnetic-Marine divergence (F_{ST})	Benthic-Limnetic divergence (F_{ST})
chrI	21,492,932	21,494,505	0.61697	0.64323	0.0019008
chrI	21,494,831	21,497,000	0.65107	0.58082	0.01914
chrI	21,499,500	21,505,000	0.53989	0.52864	0
chrI	21,514,500	21,517,500	0.45505	0.45312	0
chrI	21,527,239	21,529,500	0.60202	0.58257	0
chrI	21,531,500	21,535,598	0.55705	0.49774	0.012849
chrI	21,537,686	21,539,500	0.43466	0.39337	0.043309
chrI	21,540,000	21,545,500	0.3998	0.39948	0.037351
chrI	21,546,000	21,549,000	0.56936	0.53185	0.019607
chrI	21,552,500	21,563,000	0.39477	0.37872	0.025211
chrI	21,569,500	21,573,500	0.62772	0.56827	0.023213
chrI	21,581,500	21,586,000	0.5816	0.53054	0
chrI	21,588,500	21,593,500	0.56071	0.54237	0.019011
chrI	21,595,000	21,615,500	0.5904	0.54679	0.036269
chrI	21,619,500	21,627,500	0.58851	0.5596	0.03303
chrI	21,630,000	21,637,000	0.59911	0.5423	0.0087377
chrI	21,643,000	21,649,500	0.56538	0.5258	0.005816
chrI	21,663,000	21,669,500	0.52105	0.47171	0.033364
chrI	21,671,000	21,674,500	0.61824	0.61516	0.028671
chrI	21,675,500	21,681,500	0.39941	0.4034	0
chrI	21,683,500	21,686,500	0.36661	0.36393	0.0069198
chrI	21,694,500	21,701,000	0.46002	0.44601	0.029173
chrI	21,701,500	21,704,698	0.48712	0.45704	0.03969
chrI	21,710,087	21,710,808	0.58064	0.53083	0
chrI	21,712,479	21,713,000	0.3353	0.31988	0.032477
chrI	21,713,500	21,720,000	0.55381	0.53449	0.0046873
chrI	21,797,000	21,802,000	0.52839	0.52681	0.059318
chrI	21,803,500	21,810,000	0.44126	0.42855	0.039186
chrI	21,819,500	21,822,500	0.48191	0.44916	0.020939
chrI	21,823,283	21,827,218	0.46398	0.45642	0.00048823
chrI	21,830,000	21,836,000	0.51792	0.49199	0.00096608
chrI	21,849,000	21,855,000	0.44415	0.41288	0.015708
chrI	21,877,214	21,881,000	0.41806	0.39618	0.0039405
chrI	21,890,000	21,896,000	0.4692	0.44439	0.017035
chrI	21,899,500	21,902,000	0.48308	0.48864	0.016933
chrI	21,914,000	21,918,500	0.3741	0.37241	0.00045922
chrI	21,919,000	21,921,500	0.54714	0.42442	0.0089104
chrI	21,932,500	21,944,000	0.62745	0.59867	0
chrII	414,000	414,766	na	na	na

chrII	415,361	416,369	0.44045	0.20998	0.151
chrII	417,679	419,050	0.34186	0.11656	0.14278
chrIV	12,800,238	12,814,051	0.55915	0.49459	0.20121
chrIV	12,814,692	12,825,223	0.69007	0.50606	0.18378
chrIV	12,825,643	12,831,803	0.6481	0.42047	0.21672
chrIV	13,916,368	13,918,000	0.42739	0.26935	0.18084
chrIV	13,975,500	13,978,000	0.42739	0.37328	0.038347
chrIV	15,059,500	15,060,012	0.60932	0.18254	0.22212
chrIV	15,061,411	15,062,000	0.5643	0.17407	0.21837
chrIV	19,766,000	19,769,000	0.6899	0.66804	0.067558
chrIV	19,811,668	19,822,293	0.69003	0.69672	0.047248
chrIV	19,825,662	19,829,118	0.68659	0.65264	0.01252
chrIV	19,851,445	19,855,000	0.74825	0.77822	0.044729
chrIV	19,867,000	19,872,500	0.4195	0.44126	0.07006
chrIV	19,875,500	19,878,000	0.47172	0.45408	0.036541
chrIV	19,880,500	19,896,176	0.54676	0.55733	0.026512
chrIV	19,897,883	19,903,500	0.64914	0.66076	0.053736
chrIV	21,604,500	21,608,500	0.33015	0.035174	0.17236
chrIV	23,958,000	23,964,000	0.62162	0.23201	0.26459
chrIV	23,964,500	23,976,000	0.62436	0.20852	0.34214
chrIV	23,976,500	23,982,190	0.60571	0.13928	0.3204
chrVII	17,990,965	17,998,000	0.60223	0.083146	0.59845
chrVII	18,000,274	18,001,173	0.65325	0.087762	0.68547
chrVII	18,001,286	18,002,792	0.63457	0.075636	0.66453
chrXIX	2,452,500	2,455,500	0.6945	0.43826	0.18118
chrXIX	2,456,000	2,458,500	0.53638	0.29192	0.14766
chrXIX	2,459,000	2,471,000	0.66547	0.43967	0.15743
chrXIX	2,474,500	2,478,157	0.65659	0.42156	0.16888
chrXIX	2,488,500	2,489,124	0.42659	0.25909	0.10381
chrXIX	2,496,794	2,498,500	0.53755	0.20257	0.1239
chrXIX	2,505,500	2,507,823	0.45838	0.21391	0.095825
chrXIX	2,521,000	2,526,000	0.58646	0.20778	0.10033
chrXIX	2,545,500	2,549,000	0.46542	0.18467	0.11039
chrXIX	14,779,000	14,781,500	0.63915	0.14702	0.3111
chrXIX	14,791,535	14,794,951	0.59032	0.10608	0.3184
chrXIX	14,795,762	14,802,839	0.58578	0.10448	0.23442
chrXVIII	889,452	891,500	0.20568	0.002102	0.19389

Note: na, not available

The coordinates of adaptive regions of marine and freshwater sticklebacks were obtained from (Jones et al 2012b)

Appendix Table 14A. Correlations of genes showing *cis*-regulatory divergence between Paxton benthics and limnetics (F1 cross: BL_7)

	BL_7_1	BL_7_2	BL_7_3	BL_7_4
BL_7_1	1	0.3555616	0.3184637	0.400063
BL_7_2	0.3555616	1	0.4396796	0.5451011
BL_7_3	0.3184637	0.4396796	1	0.5257078
BL_7_4	0.400063	0.5451011	0.5257078	1

Appendix Table 14B. Correlations of genes showing *cis*-regulatory divergence between Paxton benthics and limnetics (F1 cross: BL_8)

	BL_8_1	BL_8_2	BL_8_3	BL_8_4
BL_8_1	1	0.3793044	0.4120167	0.5470489
BL_8_2	0.3793044	1	0.4902243	0.4273279
BL_8_3	0.4120167	0.4902243	1	0.4466527
BL_8_4	0.5470489	0.4273279	0.4466527	1

Appendix Table 14C. Correlations of genes showing *cis*-regulatory divergence between Paxton benthics and limnetics (F1 cross: LB_10)

	LB_10_1	LB_10_3	LB_10_4	LB_10_5
LB_10_1	1	0.4721288	0.5065735	0.4376568
LB_10_3	0.4721288	1	0.519615	0.4183301
LB_10_4	0.5065735	0.519615	1	0.4962198
LB_10_5	0.4376568	0.4183301	0.4962198	1

Appendix Table 14D. Correlations of genes showing *cis*-regulatory divergence between Paxton benthics and limnetics (F1 cross: LB_11)

	LB_11_1	LB_11_2	LB_11_3	LB_11_4
LB_11_1	1	0.4345259	0.3406198	0.3579214
LB_11_2	0.4345259	1	0.3523375	0.3721846
LB_11_3	0.3406198	0.3523375	1	0.5569055
LB_11_4	0.3579214	0.3721846	0.5569055	1

Appendix Table 15. CSS of collagen genes.

No.	Name	Chromosome	Start	Stop	minCSS
1	*COL21A1*	chrVI	7,710,406	7,724,080	**0.00145**
2	*COL14A1B*	chrXX	7,918,797	7,991,098	**0.00183**
3	*COL7A1*	chrUn	2,340,500	2,384,652	**0.00134**
4	*COL24A1*	chrVIII	7,055,749	7,103,293	0.0017
5	*COL5A2A*	chrXVI	5,805,247	5,824,442	0.00081
6	*COL23A1*	chrIV	11,443,468	11,468,190	0.00028
7	*COL11A1*	chrIII	11,565,983	11,627,261	-0.00001
8	*COL4A2*	chrXVI	7,820,982	7,852,127	0.00014
9	*COL6A4A*	chrXX	182,137	239,263	-0.00005
10	*COL5A3A*	chrXI	7,608,684	7,640,265	0.00004
11	*COL1A2*	chrX	9,261,462	9,277,295	0.00003
12	*COL28A1B*	chrXX	14,888,175	14,901,113	-0.00001
13	*COL15A1B*	chrIII	4,852,681	4,875,412	0.00001
14	*COL4A1*	chrXVI	7,880,804	7,897,351	-0.00002
15	*COL2A1*	chrXVII	5,595,183	5,627,275	-0.00001
16	*COL19A1*	chrVI	7,110,337	7,138,465	0
17	*COL1A1A*	chrXI	910,083	929,209	0.00005
18	*COL14A1A*	chrX	14,851,285	14,896,889	0.00003
19	*COL12A1B*	chrXVIII	12,414,858	12,467,049	-0.00003
20	*COL16A1*	chrX	10,630,189	10,664,614	-0.00002
21	*COL17A1B*	chrVI	10,958,264	10,974,992	0
22	*COL9A1B*	chrVI	7,095,248	7,106,403	-0.00001
23	*COL2A1A*	chrXII	1,468,616	1,485,189	-0.00005
24	*COL28A2A*	chrXVI	10,450,583	10,461,648	0.00001
25	*COL8A2*	chrX	10,166,847	10,172,167	0
26	*COL18A1*	chrXVI	4,963,171	4,986,495	0.00001
27	*COL4A3*	chrIII	2,235,826	2,255,825	-0.00005
28	*COL11A2*	chrUn	14,299,807	14,329,142	0.00001
29	*COL10A1A*	chrXV	8,673,544	8,677,560	0.00004
30	*COL27A1B*	chrXIII	17,704,841	17,747,886	-0.00004
31	*COL28A1A*	chrUn	14,138,646	14,161,350	-0.00003
32	*COL8A1A*	chrXVI	6,656,897	6,659,369	0.00005
33	*COL9A3*	chrUn	17,257,227	17,269,246	-0.00002
34	*COL12A1*	chrXV	3,469,650	3,516,066	-0.00005
35	*COL9A2*	chrX	10,666,995	10,679,875	-0.00003
36	*COL17A1A*	chrIX	2,327,057	2,339,020	-0.00001
37	*COL5A3B*	chrIX	4,485,775	4,506,442	-0.00002
38	*COL5A1*	chrXIV	2,012,587	2,051,092	-0.00003
39	*COL6A3*	chrXVI	2,063,216	2,097,073	-0.00001
40	*COL10A1B*	chrXVIII	3,331,871	3,333,721	0.00004