Combining genomics and transcriptomics to study adaptation to lake and river habitats in three-spined sticklebacks

Dissertation in fulfilment of the requirements for the degree

Doctor rerum naturalium

of the Faculty of Mathematics and Natural Sciences

at Kiel University

submitted by

Yun Huang

Plön, 2017

First referee:	Prof. Dr. Manfred Milinski
Second referee:	Prof. Dr. Tal Dagan
Date of the oral examination:	30.01.2018
Aprroved for publication:	31.01.2018

Signed: Dean

Table of Contents

Summary

Understanding the genetic basis of adaptive evolution is a prime objective in modern evolutionary studies. However, disentangling adaptive and neutral evolution remains a challenging task. Parallel evolution, where similar phenotypes independently arise in similar environments, provides compelling evidence for adaptation, as the repeated emergence of similar phenotypes is unlikely to happen due to neutral processes alone. The three-spined stickleback (*Gasterosteus aculeatus*) represents an ideal system to study parallel evolution due to its rapid adaptation to various freshwater habitats since the last glaciation. The repeated adaptation to lake and river habitats has been proposed to be driven by distinct parasite environments. This has resulted into distinct lake and river ecotypes differing in their parasite defense. In this thesis, I investigated the magnitude of genetic parallelism and habitat-specific gene expression underlying the repeated phenotypic adaptation to the distinct habitats of lakes and rivers. In my first chapter I developed a novel genome scan approach based on mutual information criteria. By applying this approach to whole-genome sequencing data of wild-caught three-spined sticklebacks from five parapatric lake river population pairs, I detected a low degree of parallel genetic changes across these geographically widespread population pairs. In contrast, in my second chapter, transcriptome profiling of two immune tissues from a subset of the individuals used for the genome study discovered habitat-specific gene expression patterns. Such habitat-specific patterns display similar expression among the same ecotypes but different expression between ecotypes, indicating parallelism at the expression level. I identified a total of 139 genes with habitat-specific expression patterns, eight of which were annotated with immune functions and 42 differentially expressed in previous parasite exposure experiments, suggestive of a parasite defense function in nature. Integrating the genome and transcriptome analyses from the first two chapters, the last chapter addressed the genetic basis of habitat-specific gene expression. Using genome and transcriptome data from the same individual fish, I evaluated the extent of sequence divergence in cis-regulatory regions and gene copy number divergence associated with expression divergence. Though weak correlations were found genome-wide, two

genes showed significant divergence in both gene copy number and gene expression; the strong correlation between gene copy number and expression level in these two genes suggest a dosage effect impacts habitat-specific gene expression. Taken together, this thesis provides a detailed view on genetic and transcription divergence between lake and river sticklebacks, and describes the complex and idiosyncratic nature of evolution at the genetic level. My contributions support the idea that gene expression promotes repeated adaptation to lake and river environments, largely influenced by non-parallel mutations, but in some cases facilitated by recurrent copy number changes at the genetic level.

Zusammenfassung

Ein Verständnis der genetischen Grundlagen von adaptiver Evolution ist eines der wichtigsten Ziele evolutionsbiologischer Studien. Allerdings besteht die Schwierigkeit adaptive und neutrale Evolution auseinanderzuhalten. Parallele Evolution, d.h. das unabhängige Auftreten gleichartiger Phänotypen in analogen Lebensräumen, stellt ein starkes Argument für Anpassung dar, da ein wiederholtes Entstehen gleichartiger Phänotypen aufgrund von neutralen Prozessen allein unwahrscheinlich ist. Nach der letzten Eiszeit, hat sich der Dreistachlige Stichling (*Gasterosteus aculeatus*) mehrfach an ein Leben in verschiedenen Süßwasser-Lebensräumen angepasst und bietet damit die idealen Voraussetzungen parallele Evolution zu erforschen. Es wurde vorgeschlagen, dass diese wiederholte Anpassung an See- und Fluss -Lebensräume maßgeblich durch Selektion aufgrund von lebensraumspezifischer Parasiten vorangetrieben worden sein. Dies führte auch dazu, dass sich See- und Fluss-Phänotypen in ihrer Verteidigung gegen Parasiten unterscheiden. In dieser Doktorarbeit untersuchte ich den Umfang genetischer Parallelität und lebensraumspezifischer Genexpression, welche den analogen Phänotypen dieser Lebensräume zu Grunde liegt. In meinem ersten Kapitel entwickelte ich einen neuartigen Genom-Scan-Ansatz, der auf dem Kriterium der „Mutual Information“ basiert. Durch die Anwendung dieses Ansatzes auf Sequenzierungsdaten ganzer Genome bestätigte ich ein niedriges Ausmaß an parallelen genetischen Veränderungen. Im Gegensatz dazu entdeckte ich in meinem zweiten Kapitel durch „Transcriptome-Profiling“ von zwei Immun-Geweben, lebensraumspezifische Genexpressionmuster. Es zeigt sich, dass die Expressionsmuster gleichartiger Phänotypen ähnlich sind, während diese sich zwischen See- und Fluss-Phänotypen unterscheiden. Diese Ergebnisse deuten auf Parallelität auf der Expressionsebene hin. Insgesamt identifizierte ich 139 Gene mit solchen lebensraumspezifischen Expressionsmustern, von denen acht mit Immunfunktionen in Verbindung gebracht werden können. Darüberhinaus, fand ich 42 unterschiedlich exprimierte Gene, welche in früheren Infektionsexperimenten, bereits identifiziert wurde, da diese Gene eine Rolle bei der Parasitenverteidigung im natürlichen Lebensraum spielen. Durch die

Kombination von Genom- und Transkriptom-Daten von den gleichen Fischindividuen, evaluierte ich in meinem letzten Kapital in welchem Maß Sequenzunterschieden in cis-regulatorischen Regionen und Unterschiede in der Gen-Kopie-Anzahl mit der Divergenz in der Genexpression assoziiert sind.

Obwohl nur schwache Korrelationen Genome-weit gefunden wurden, zeigten besonders zwei Gene einen signifikanten Unterschied sowohl in der Gen-Kopie-Anzahl als auch bei der Genexpression. Die korrelierten Gen-Kopie-Anzahlen und Expressionsniveaus dieser beiden Gene legen eine Dosiswirkung nahe, welche die lebensraumspezifischen Genexpression dieser Gene beeinflusst. Zusammengefasst, diese Arbeit liefert eine detaillierte Sicht auf die genomischen und transkriptomischen Unterschiede zwischen See- und Fluss-Stichlingen und offenbart die komplexe Natur der Evolution auf genetischer Ebene. Meine Arbeit unterstützt die Ansicht, dass Genexpression die wiederholte Anpassung an spezifische Lebensräume fördert, zum größten Teil auf der genetischen Ebene verursacht durch spezifische, nicht parallele Mutationen, aber in machen Fällen auch durch sich wiederholende Änderungen in der Gen-Kopie-Anzahl.

Introduction

Adaptation, the process by which a population of organisms evolves fitter forms to survive in the environment due to natural selection, is a central paradigm in evolutionary biology. Finding the genetic changes and molecular modifications between populations that facilitate adaptation to different environments helps to understand the origin and targets of adaptive evolution. In this general introduction to my PhD thesis, I will elaborate on the signatures of adaptation on genomes and transcriptomes, and how to identify these signatures utilizing cases of parallel evolution.

1. Adaptive evolution and parallel evolution

The theory of adaptation by natural selection comprises a centerpiece in the subject of evolution (Darwin 1859; Fisher 1930). Natural selection can occur when heritable traits vary in form and fitness in a population. Adaptive traits that offer fitness advantages tend to increase in frequency over time. Though adaptation by natural selection is a ubiquitous phenomenon in nature (Endler 1986), evolution would not necessarily be adaptive. Neutral evolution differs from adaptive evolution in that neutrally evolving traits are irrelevant to fitness and the changes of trait frequencies are mainly due to random effects (Kimura 1983). Neutral processes such as genetic drift, migration and demographic history also influence the evolutionary trajectories of traits and can lead to differentiation between populations (Kimura 1968). For example, genetic drift alone can generate significant changes of frequencies for neutral traits and even slightly deleterious traits, especially in populations of small sizes (Charlesworth 2009). When the fitness value of a trait is unknown, a significant increase in its frequency can thus be mistaken as a signal of adaptive evolution. Disentangling adaptive evolution from neutral evolution remains challenging in evolutionary studies aiming to understand the importance of natural selection in evolution.

Habitat plays an important role in the action of natural selection, as it harbors essential resources for organisms to survive and reproduce and therefore impacts fitness. Habitats differing in abiotic and biotic resources can lead to diversification of populations when traits are differentially favoured in a given habitat (Nosil & Feder 2012; Savolanen et al 2013). In contrast, ecologically similar habitats from geographically distant locations might independently favour similar traits of the inhabiting populations (Elmer & Meyer 2011), giving rise to ecotypes or ecomorphs (Turesson 1922; Turrill 1946). The repeated emergence of similar traits in independent lineages is termed parallel evolution (Futuyma 1986). Under neutrality, populations or species are expected to diverge along with the phylogenetic relationships (Kreitman 1996; Orr 1998), which means traits are more similar among closely related taxa than distantly related taxa. Deviations from this expectation, such as parallel evolution seen in ecotypes, provide strong evidence for natural selection (Endler 1986, Figure 1).

Parallel phenotypic evolution in the wild has been widely documented. Examples include independent eye reduction and antennae elaboration in cave animals (such as amphipods *Gammarus minus* by Jones et al. 1992); independent origins of *Anolis* lizards ecomorphs among Caribbean islands (Losos et al. 1998); parallel

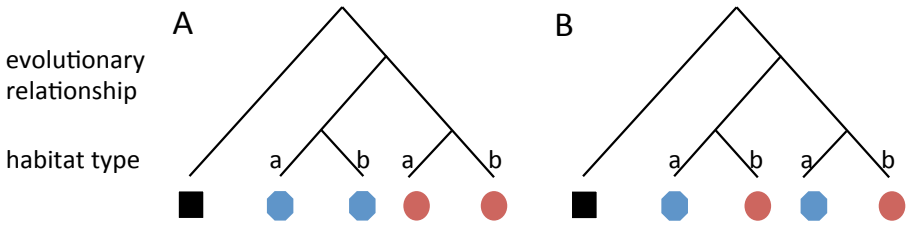


Figure 1. Scenarios of (A) a possible neutral scenario where traits cluster by the phylogenetic relationship versus (B) an adaptive scenario where traits correlate with habitat types in different clades. The dendrograms indicate the phylogenetic relationship of the taxa inhabiting different habitat types (a and b). Black squares indicate outgroups, while other colours and shapes indicate different traits. Modified from Figure 1, Whitehead 2011.

differentiation of dwarf and normal whitefish in North American lakes (Bernatchez & Dodson 1991; Pigeon et al. 1997); parallel increase of wing length with latitude in both the New World and the Old World in *Drosophila subobscura* (Huey et al. 2000); repeated differentiation of pharyngeal jaw and lip shapes in cichlids (Albertson et al. 2005); repeated differences of body depth and gill raker numbers between lake and stream sticklebacks (Berner et al. 2008; Kaeuffer et al. 2012) and enlarged head sizes of geckos in multiple island populations (Eloy de Amorim et al. 2017). Most cases of parallel evolution have been described based on morphological or physiological phenotypes. The underlying molecular basis of parallel evolution, such as genetic modifications that affect protein function or gene expression, are comparatively less well understood. With recent advances in sequencing technologies, we are able to tackle these issues.

2. Adaptive evolution of gene expression

In general, two types of molecular changes lead to phenotypic effects. One involves changes in protein structure that alter physical and chemical properties, changing activities and molecular functions of the protein (Hoekstra & Coyne 2007). Another type involves regulatory changes that do not alter protein structure, but rather the amount and/or spatiotemporal expression of gene products (King & Wilson 1975). Regulatory changes have been proposed to play a more important role in adaptive evolution (Carroll 2008), as gene expression is dynamic and flexible whereas protein structures are comparatively pleiotropic and constrained (Wray 2007). Prominent examples of gene expression changes contributing to adaptive phenotypic evolution include calmodulin expression changes affecting beak morphology of Darwin's finches (Abzhanov et al. 2006), increased expression of Agouti producing camouflage pigmentation in deer mice (Linnen et al. 2009), and the silencing of expression of *pitx1* associated with repeated pelvic loss in three-spined sticklebacks (Chan et al. 2010).

Gene expression itself can be considered as an extended molecular phenotype, representing the molecular basis for morphological or physiological phenotypes (Houle et al. 2010). Due to the ease and accessibility of gene expression data

from microarrays and RNA-seq technologies, transcription is commonly used as a proxy for gene expression and studied genome-wide (Schena et al. 1995; Alvarez et al. 2015). With notable degrees of heritability (Stamatoyannopoulos 2004; Gibson & Weir 2005) and contribution to fitness (Pavey et al. 2010), variation in gene expression is a substrate for natural selection (Figure 2A). However, what proportion of gene expression is under adaptive evolution remain elusive. Some studies suggest that the expression of most genes are under neutral evolution (Oleksiak et al. 2002; Khaitovich et al. 2005), while other studies suggest that a large component of gene expression evolution is under either stabilizing selection (Rifkin et al. 2003; Lemos et al. 2005) or directional selection (Enard et al. 2002; Fraser et al. 2010; Nourmohammad et al. 2017). Using comparative transcriptomic approaches, the selection regimes of gene expression evolution can be inferred based on expression differences between populations compared to that of within populations (Harrison et al. 2012). This analysis can be extended towards studying gene expression patterns across populations correlated with particular environments. For example, gene expression variation was found to be associated with a gradient of habitat temperatures in killifish (*Fundulus heteroclitus*), suggesting local adaptation (Whitehead & Crawford 2006). When contrasting discrete habitats, approaches aiming to identify the genes differentially expressed between habitats are most powerful if habitat contrasts are replicated samples (Figure 2B). This allows identifying habitat-specific expression pattern, which is a manifest of parallelism at the expression level. Studies of this kind have revealed parallel expression patterns between sympatric 'dwarf' and 'normal' whitefish ecotypes specializing in limnetic and benthic niches (Derome et al. 2006), between lake and river ecotypes of salmon (Pavey et al. 2011), between thick- and thin-lipped Midas cichlid ecomorphs (Manousaki et al. 2013), and between low and high latitude populations of *Drosophila melanogaster* (Zhao et al. 2015). Identifying those genes differing in their expression in parallel is one step forward to understand the molecular targets of natural selection. However, which proportion of genes shows an expression difference due to plastic versus genetic differences remains

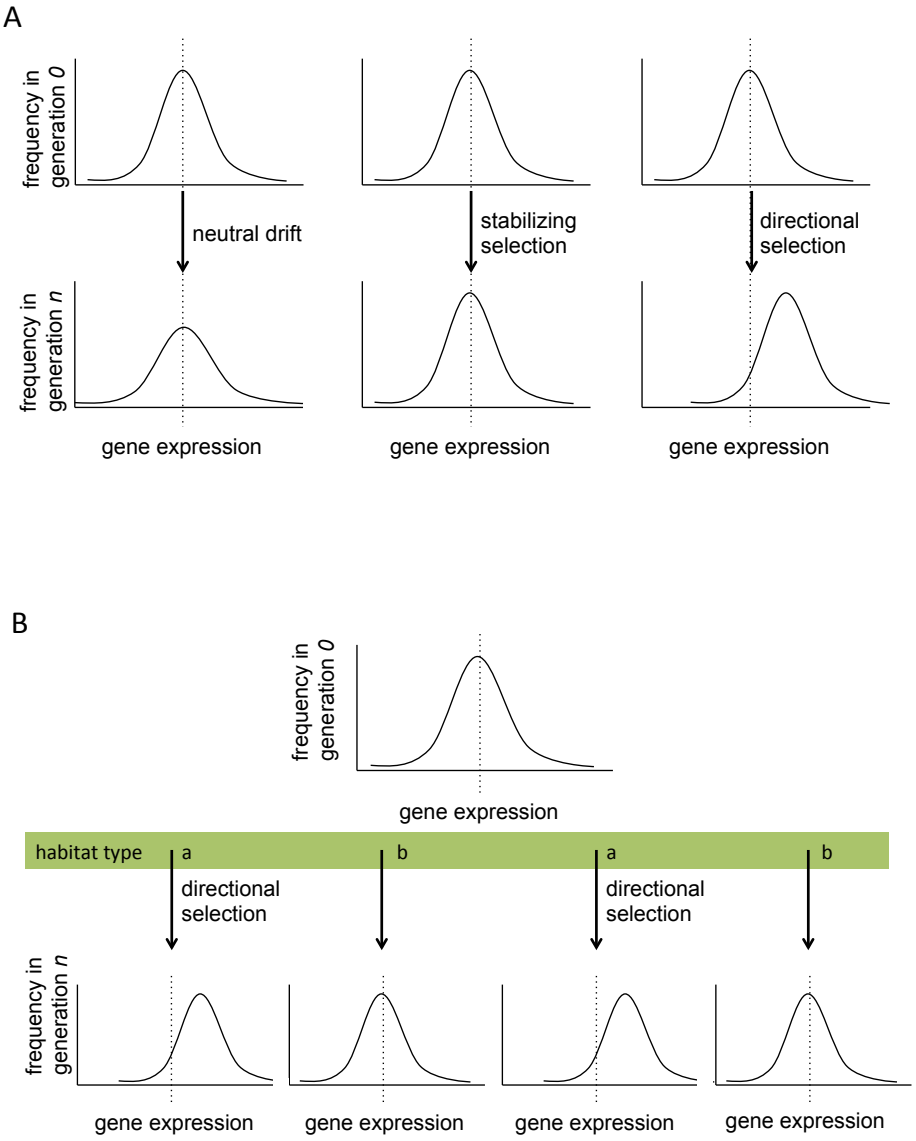


Figure 2. Schematic illustration of gene expression evolution given different selection scenarios. (A) Changes in the frequency distribution of gene expression levels between an ancestral population (upper row) and descendant populations (bottom row) under models of neutral drift, stabilizing selection and directional selection. The dotted vertical line represents the mean expression in the ancestral population. Neutral drift flattens the distribution, stabilizing selection maintains the population mean while directional selection shifts the population mean to higher or lower levels. (B) Parallel gene expression evolution where population means shift independently in the same way in similar habitats (habitat type *a* in this example). This is expected under habitat-associated directional selection, governing gene expression evolution in either or both contrasting habitats.

challenge to disentangle. A plastic response can play a beneficial role when populations are exposed to a new environment, but its evolutionary significance remains controversial (Laland et al. 2014). A heritable basis of expression differences provides evidence for adaptive evolution that the underlying genetic variants were differentially selected upon by different environments. Hence, unraveling the underlying genetic basis of gene expression differences is important to understand the targets of adaptive expression evolution.

3. Genetics of adaptive expression evolution

Genetically based gene expression differences between populations or species are mainly attributed to two types of genetic variations: sequence variation, most commonly identified by single nucleotide polymorphisms (SNPs), and structural variation (SV), which includes copy number variations (CNVs) such as duplications and deletions of large genome regions. These two types of variation are predicted to affect gene expression through different ways of regulation.

Sequence variation that occurs in regulatory regions can alter gene expression. Regulatory regions are genomic regions involved in launching transcription and are classified as *cis*- and *trans*-regulatory regions. *Cis*-regulatory regions (or *cis*-regulatory elements, CREs, Figure 3A) contain transcription factor binding sites that are necessary to initiate transcription, including promoters and enhancers, located in the neighborhood of the transcribed genes. *Trans*-regulatory regions (or *trans*-regulatory element, TREs, Figure 3B) are remote genomic regions that typically encode transcription factors that bind to CREs. Sequence changes both in CREs and in TREs can alter the binding affinity of transcription factors and thus affect expression of adjacent or remote genes. Consistent with the rationale that CREs have mostly local effects on gene expression compared to TREs and thus lower pleiotropy, CREs are favored during population divergence and are primary drivers of gene expression differences between species (Wittkopp et al. 2008; Emerson et al. 2010). There is increasing evidence that CRE sequence divergence is important in phenotypic evolution; expression quantitative trait loci (eQTL) studies so far have found

predominately *cis*- compared to *trans*-localized eQTL (Dixon et al. 2007; Stranger et al. 2007a; Bryois et al. 2014; Pritchard et al. 2017; Ishikawa et al. 2017). Examples of the importance of CREs in gene expression changes include nucleotide substitutions within CREs altering cis-regulatory activities and responsible for gene expression differences and loss of trichomes in Drosophila species (Frankel et al. 2011), and 13 nucleotide substitutions in an enhancer responsible for human-specific limb development compared to chimpanzees (Prabhakar et al. 2008).

Structural variation can also influence gene expression, for example by changing the number of gene copies via copy number variation. Such CNV genes typically show dosage effects on gene expression, where higher gene copy numbers lead to higher gene expression, simply because more copies are available for transcription, and vice versa (Figure 3C; Haraksingh & Snyder 2013; Gamazon & Stranger 2015). Other factors can obscure a dosage effect of CNVs, such as compensatory effects via negative feedback in regulation networks or differences in chromatin profiles among copies (Henrichsen et al. 2009a). Nevertheless, genome-wide association studies and CNV surveys have suggested a great impact of CNVs on gene expression (Sudmant et al. 2015; Huddleston & Eichler 2016). A substantial proportion of genes within CNV regions or in their vicinity have expression levels correlated with copy number, as has been reported in Drosophila, mice and human (Stranger et al. 2007b; Schlattl et al. 2011, Henrichsen et al. 2009b, Cardoso-Moreira et al. 2016).

Because gene expression can be modulated by different genetic variants, it raises the possibility that parallel gene expression patterns can be caused by entirely different types of mutations. It is therefore an informative question to ask whether parallel gene expression changes are accompanied by parallel genetic changes, or whether the underlying genetic causes involve different nucleotide sites, different genes, or different mutation types. In addition, adaptive genetic changes can arise via new mutation, from standing genetic variation, or though introgression (Elmer & Meyer 2011; Stern 2013). Adaptation from standing genetic variation in shared ancestral populations might be a rapid trajectory for

adaptive evolution compared to new mutations, as it does not require waiting time for mutations to arise and integrate into the existing functional networks (Barrett & Schluter 2008). Concluding from experimental studies, population genomic studies and quantitative genetic studies, parallel adaptation at the genetic level appears to be frequent, especially for traits with a simple genetic basis of large effect (Wood et al. 2005), and when populations share recent common ancestors (Conte et al. 2012). But how much genetic parallelism is underlying expression parallelism on a genome wide scale is unclear.

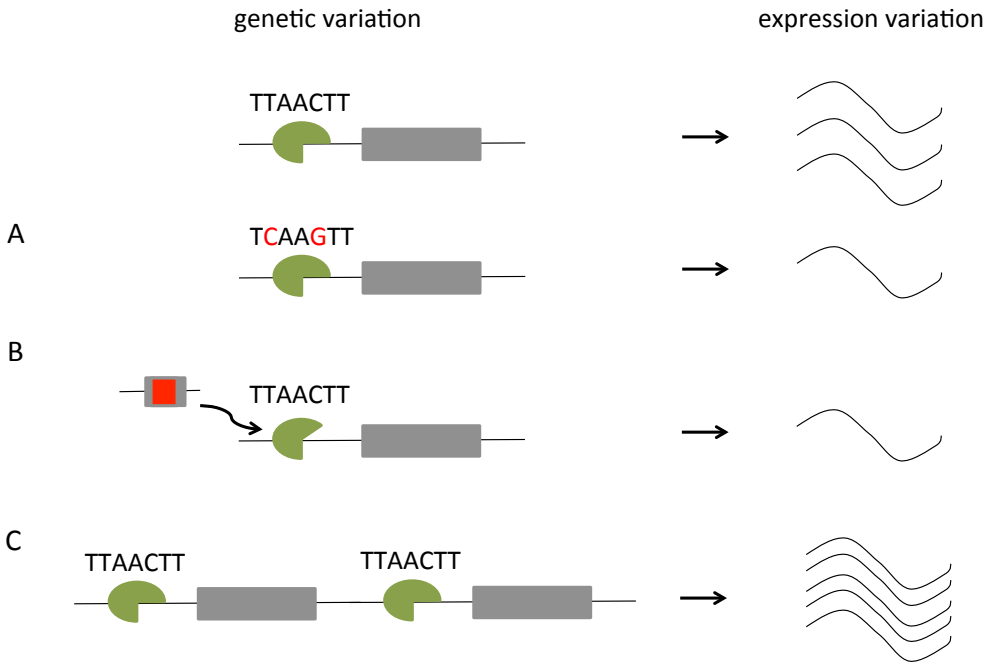


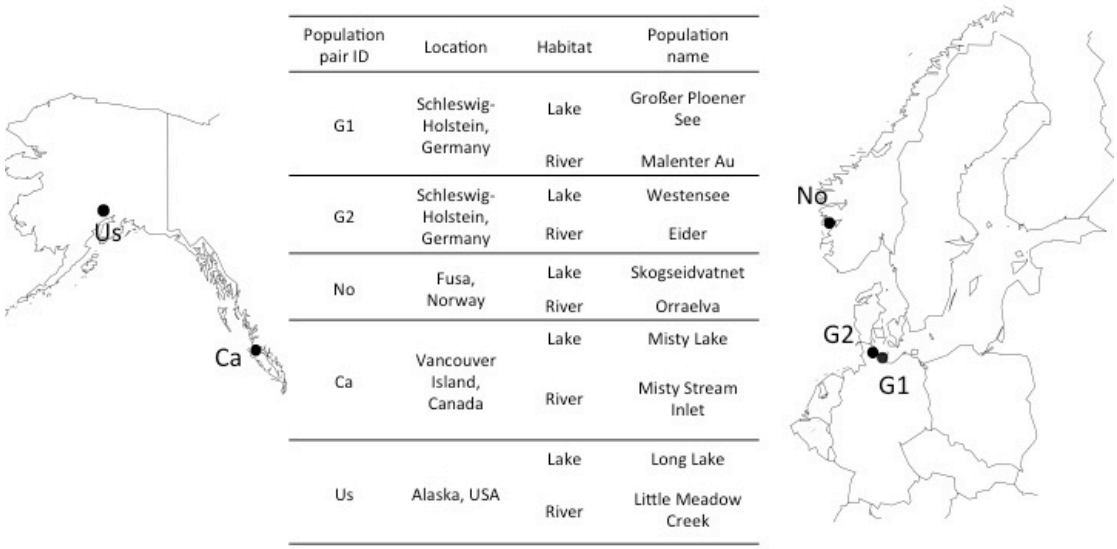
Figure 3. Genetic variation affecting gene expression levels. Green pies indicate CRE regions where transcription factors bind, and grey boxes indicate gene bodies. (A) Nucleotides in red indicate single nucleotide polymorphisms in CRE regions that alter transcription levels. (B) The additional box with a red part indicates mutations in TRE regions that alter transcription levels. (C) Higher gene copy number can lead to higher transcription levels.

4. The three-spined stickleback

The three-spined stickleback (*Gasterosteus aculeatus*) is a great model organism to study the genetic and molecular basis of parallel evolution, due to its rapid radiation in different habitats and broad distribution in the northern hemisphere. After the last glaciation, marine sticklebacks rapidly and repeatedly colonized different freshwater habitats, giving rise to parallel freshwater ecotypes (McKinnon & Rundle 2002). There are numerous examples of parallel genetic changes underlying the adaptation to freshwater habitats, including the repeated loss of the pelvic fin caused by independent deletion events frequently occurring in the CRE regions of *pitx1* gene that suppress gene expression (Chan et al. 2010), and the loss of armor plates in multiple freshwater populations via repeated fixation of the same Ectodysplasin alleles (Colosimo et al. 2005; but see Pujolar et al. 2017). Genome-wide surveys of parallel evolution have suggested that adaptation of sticklebacks to freshwater habitats predominantly involves regulatory changes compared to protein-coding changes (Jones et al. 2012). Amongst various freshwater habitats, recurrent adaptation to lakes and rivers has given rise to two distinct ecotypes (Reusch, et al. 2001). Adjacent lake and river populations often exhibit consistent morphological differences in body shape and gill raker number (Kaeuffer et al. 2012; Lucek et al. 2014). Amongst other differences between lake and river habitats, one profound difference imposing natural selection on the organism is higher parasite abundance in lakes fish than in river fish (Kalbe et al. 2002). Lake fish commonly bear a higher parasite load than the parapatric river populations (Eizaguirre et al. 2011), resulting in higher immune-competence (Scharsack et al. 2007) and higher diversity in major histocompatibility complex (MHC) (Eizaguirre et al. 2011). It appears as though this emerges due to adaptation to the local parasite environments (Eizaguirre et al. 2009, 2012b). Altogether, these studies suggest that the differences in habitats such as the parasite pressure play an important role in the divergence between lake and river sticklebacks.

In this thesis, wild-sampled sticklebacks from five pairs of parapatric lake and river populations are used to investigate parallelism on the genetic level as well

as on the gene expression level (Figure 4). Studying the genomic and transcriptomic variation between these ecotypes allows a better understanding of the genetic and molecular basis underpinning the habitat-specific adaptation in lake and river sticklebacks.



Population pair ID	Location	Habitat	Population name
G1	Schleswig-Holstein, Germany	Lake	Großer Ploener See
		River	Malenter Au
G2	Schleswig-Holstein, Germany	Lake	Westensee
		River	Eider
No	Fusa, Norway	Lake	Skogseidvatnet
		River	Orraelva
Ca	Vancouver Island, Canada	Lake	Misty Lake
		River	Misty Stream Inlet
Us	Alaska, USA	Lake	Long Lake
		River	Little Meadow Creek

Figure 4. Sampling sites of lake and river population pairs included in this thesis. Modified from Figure 1 in Feulner et al. 2015.

Thesis outline

The aim of my PhD study was to understand the genetic basis of habitat-specific adaptation, for which I studied variation in genomes and transcriptomes among five replicated lake and river population pairs of three-spined stickleback. My PhD study comprises three projects, which are presented as three chapters in this thesis and outlined below. These projects were conducted in cooperation with other scientists, with their and my contribution indicated in a table at the end of this section.

Chapter I. Genomic parallelism detection via mutual information

A genome scan based on mutual information (MI) from information theory was employed to evaluate parallel sequence divergence between lake and river sticklebacks across five population pairs. This was carried out in two complementary ways: (a) contrasting each parapatric lake and river populations, as it was done by F_{ST}, allowing the identification of genomic regions with exceptionally high divergence that were shared between population pairs; (b) contrasting grouped lake or river populations altogether to evaluate the sequence divergence between ecotypes from all sampled population pairs. The latter was not performed on F_{ST} because F_{ST} is based on the assumption of interbreeding systems. These MI-based analyses confirmed with the F_{ST} analyses that little genetic parallelism was found across the replicated population pairs.

Chapter II. Habitat-specific gene expression in sticklebacks

Parapatric lake and river populations of sticklebacks are known to harbor distinct parasite communities, a parallel difference across multiple lake and river pairs driving divergence between lake and river populations. Such parasite-mediated selection should leave habitat-specific signatures on gene expression divergence especially in immune tissues and genes with an immune function. These signatures should exhibit similar expression levels within ecotypes compared to between ecotypes. Using a subset of the stickleback

samples used for the genome study from the first chapter, two major immune tissues were used for whole-transcriptome sequencing. Differential gene expression analysis revealed 139 genes with habitat-specific expression patterns. Amongst these genes, eight were annotated with immune functions and 42 were found differentially expressed in previous parasite-infection studies, suggestive of a contribution in coping with differential parasite pressures between lake and river habitats.

Chapter III. Genetics underlying habitat-specific gene expression

This chapter aimed to understand the genetic basis of habitat-specific gene expression. Sequence divergence in *cis*-regulatory regions, gene copy number divergence and expression divergence between ecotypes were evaluated by ANOVA-based F-statistics and integrated to examine the contribution of genetic divergence in gene expression divergence. Weak positive correlations were found genome-wide. However, two genes with expression divergence between ecotypes exhibited highly correlated copy number divergence, while no such case was found for *cis*-sequence divergence in our data. These two genes have putative immune functions and are strong candidates involved in adaptation to differential parasite environments between lake and river habitats. These results suggest that gene copy number changes can lead to gene expression changes between ecotypes, providing a putative mechanism facilitating adaptation to different habitats.

Taken together, this thesis addressed parallelism on the genetic level and gene expression level across replicated lake and river stickleback population pairs. Though little genetic parallelism was found across the replicated population pairs, 139 genes displayed habitat-specific expression patterns. In addition, we found compelling evidence that gene copy number changes of at least two genes were strongly associated with habitat specific expression patterns. These results highlight a contribution of copy number variation in adaptive evolution, advancing our understanding of the genetic basis underlying habitat-specific adaptation.

Table of contributions:

	Chapter 1	Chapter 2	Chapter 3
Project Initiation	EBB, MM, TBHR	The Big Screen Consortium	The Big Screen Consortium
Sample collections and preparations	FJJC, CE, PGDF, MK	FJJC, CE, PGDF, MK, IES, MS	-
Analyses Design	FJJC, PGDF, YH	FJJC, PGDF, YH, MP	FJJC, PGDF, YH
Analyses Performance	YH	YH	YH
Interpretation and Writing	FJJC, PGDF, YH	FJJC, PGDF, YH	FJJC, PGDF, YH

Author names are given in an alphabetical order: EBB: Erich Bornberg-Bauer; FJJC: Frédéric J. J. Chain; CE: Christophe Eizaguirre; PGDF: Philine G. D. Feulner; YH: Yun Huang; MK: Martin Kalbe; MM: Manfred Milinski; TBHR: Thorsten B. H. Reusch; IES: Irene E. Samonte; MS: Monike Stoll.

The Big Screen Consortium includes: Manfred Milinski, Thorsten B. H. Reusch, Erich Bornberg-Bauer, Monika Stoll, Martin Kalbe, Christophe Eizaguirre, Tobias L. Lenz, Philine G. D. Feulner, Frédéric J. J. Chain, Mahesh Panchal, Irene E. Samonte.

Chapter I Using mutual information to detect genomic regions of parallel differentiation between replicated parapatric lake and stream stickleback populations

Yun Huang, Philine G. P. Feulner & Frédéric J. J. Chain

Introduction

When populations split and adapt to different habitats, genotypes that are newly beneficial to a particular habitat are expected to thrive locally but diverge from other populations (Lewontin & Krakauer, 1973). Accordingly, a locus can become highly differentiated due to positive selection in one population. However, neutral evolutionary processes such as genetic drift, demographic history and reduced gene flow can also lead to high genetic differentiation (Kimura 1968). Parallel evolution, where the same genotypes repeatedly arise in high frequency in populations in similar environments and differentiate between populations in distinct environments provides strong evidence for natural selection (Elmer & Meyer 2011), as the odds of neutral processes repeatedly yielding high differentiation in the same direction is small. The study of parallel evolution in populations inhabiting different environments can help identify targets of selection and adaptation.

Genetic differentiation can be estimated as F_{ST} (Wright 1951), a function of the average genetic variance between populations compared to total genetic variance across populations. F_{ST} is commonly applied to pair-wise population comparisons, while F_{ST} in a hierarchical island model was developed to apply to multiple populations with a hierarchical structure (Slatkin & Voelm, 1991).

This chapter serves as a supplementary analysis for
Feulner et al. 2015. *PLoS Genetics* 11:e1004966.

Either pair-wise or hierarchical estimation is theoretically based on estimates of inbreeding coefficients between subpopulations, and thus not suitable to be applied to comparative approaches that violate the real population structure.

Alternatively to F_{ST}, mutual information (MI) provides a robust measure of genetic differentiation between populations based on the amount of uncertainty in the distribution of genotypes amongst populations (Dewar et al. 2011) and is free of assumptions about the underlying population structure and relatedness. In information theory, the amount of uncertainty within a random variable is termed entropy (H). It is often used to measure species diversity in ecology, also referred to as the Shannon index, which combines the number of species detected and the frequency of each (Hill 1973). In ecology, the prevalence of different species and the more even their frequencies, the higher the Shannon index. Similarly, in the context of genetic differentiation, entropy describes the evenness of the distribution of different genotypes across individuals. When dealing with populations from different habitats, the reduction of entropy in genotypes given a habitat (population) represents the MI between the genotypes and the habitat. For example, if the genotype is completely unrelated to the habitat type, then the uncertainty about the genotypes remains the same with or without the awareness of the habitat type and the MI is zero. On the contrary, if the uncertainty is reduced within a habitat compared to pooling genotypes across habitats, the MI is positive. At best, a fixed genotype between two habitats would accurately inform the habitat type, and the MI would be the maximum entropy of the habitat and normalized to one. As MI is not based on any population genetic assumptions, it can be applied to any hierarchy of population structures. This is particularly useful for studying the parallel evolution of "ecotypes". Ecotypes are ecologically divergent populations with distinct phenotypes associated with the habitat environments instead of phylogeographical distances (Turrill 1946). Because the comparison between different ecotypes does not conform to the phylogenetic relationship among populations, MI is well suited to study ecotypes and the underlying parallel genetic changes.

Three-spined sticklebacks have repeatedly colonized various freshwater habitats and have differentiated likely due to rapid adaptation since the last glaciation (McKinnon & Rundle 2002). Amongst the emerged ecotypes, lake and river sticklebacks display some parallel traits such as body depth and gill raker numbers (Berner et al. 2008; Kaeuffer et al. 2012) and parasite resistance (Eizaguirre et al. 2011; Scharsack et al. 2007), suggesting the action of natural selection. However, the genetic parallelism potentially underlying such phenotypic parallelism is not well understood.

In this chapter, I aim to study genetic parallelism across lake and river stickleback ecotypes from multiple geographically widespread population pairs. Using a MI-based approach, we scanned the whole-genome for divergence patterns. We specifically looked for genomic regions having elevated divergence (high MI) in multiple population pairs and addressed the presence and extent of genetic parallelism in stickleback ecotypes.

Methods

Genotype materials

Three-spined sticklebacks were sampled from five lake and river population pairs, two from Germany (G1 and G2), one from Norway (No), one from Canada (Ca) and one from America (Us). Six random individuals from each population were used for DNA extraction and whole-genome sequencing. The sequencing data is publicly available in the European Nucleotide Archive (PRJEB5198). Detailed information on sequence data processing was described in Feulner et al. (2015). Briefly, adapters were removed, reads with low quality base calls and PCR duplicates were filtered out. Regions of the genome that are highly repetitive (masked regions) were excluded to reduce the impact of false variant calls. A total of 7920208 reliably genotyped sites of single nucleotide polymorphisms (SNPs), which were also used in Feulner et al. (2015), were used for this MI analysis.

MI analyses

Our sampled lake and river populations (see Figure 4 in the Introduction) consisted of several hierarchical levels for which we could assess ecotype differences. The MI estimates were calculated in a total of eight comparisons, consisting of (1) within each parapatric pair (five pair-wise comparisons: G1, G2, No, Ca and Us), (2) within continents (two continent-wise comparisons: Europe and North America), and (3) across all populations together (global). For each of the comparison, the entropy of ecotype H(E) was calculated as:

$$H(E) = -\frac{r}{n} log_2 \frac{r}{n} - \frac{l}{n} log_2 \frac{l}{n}$$

where n is the total number of sampled fish, and l and r are the numbers of sampled fish from lake and river ecotypes, respectively. Because we always have balanced sample sizes between lake and river ecotypes (l=r), for all our comparisons, H(E)=1.

Analogously, the entropy of alleles H(A) was calculated.

$$H(A) = -\sum_{i=1}^{m} a_i \, log_2 \, a_i$$

where the i^{th} allele has a frequency a_i within the pool of lake and river samples, and m represents the number of different alleles occurring in the SNP site. As only bi-allelic SNP data were used, m is always 2 in our analyses.

The conditional entropy of alleles given ecotypes, lake and river respectively, are calculated as:

$$H(A|E = \text{"lake"}) = -\sum_{i=1}^{m} l_i \, log_2 \, l_i$$

$$H(A|E = \text{"river"}) = -\sum_{i=1}^{m} r_i \, log_2 \, r_i$$

where r_i and l_i represent the allele frequencies of the i^{th} allele in lake and river population pairs, respectively.

Weighted by respective sample sizes of the ecotypes, the total conditional entropy of alleles given ecotypes is:

$$H(A|E) = \Pr(E = \text{"lake"}) \; H(A|E = \text{"lake"}) + \Pr(E = \text{"river"}) \; H(A|E = \text{"river"})$$

where

$$\Pr(E = \text{"lake"}) = \frac{l}{n}$$

$$\Pr(E = \text{"river"}) = \frac{r}{n}$$

MI is quantified as the difference between the entropy of alleles and the conditional entropy of alleles given ecotypes:

$$MI(A, E) = H(A) - H(A|E)$$

MI can maximally be H(E), when the distribution of alleles exactly mirrors ecotypes, then H(A) equals H(E) and the uncertainty of alleles given ecotypes is zero (H(A|E)=0). MI can minimally be zero when H(A) equals H(A|E), meaning ecotypes does not provide information about distribution of alleles at all. Thus, MI can be normalized to MI'(A, E)=MI(A, E)/H(E), which yields values within the range from 0 to 1.

For each comparison, SNPs with minor allele frequencies below 5% were excluded to avoid biases by uninformative polymorphsms (following Roesti et al. 2012a). MI was calculated for individual SNPs, and averaged across non-overlapping 100kb windows to identify genomic regions that harbor potentially linked SNPs with elevated MI values. Outlier windows were identified based on two criteria as the following. First, we performed a permutation analysis by shuffling individuals (1000 times) among ecotypes. A window was considered as significantly differentiated if the observed MI window average was greater than 95% of the window values from permutations. Second, windows with significant MI that fell within the top 0.5% of empirical distribution were identified as outlier windows. Overlaps between different sets of outlier windows from the 8 different comparisons were identified. To compare outlier windows with the results from an F_{ST} analysis on these same data using non-overlapping 10kb windows (Feulner et al. 2015), we also calculated average MI for non-overlapping 10kb windows. Gene annotations within the overlapped outlier windows were used for functional enrichment analyses against the genome background. The calculation of MI for SNPs and permutations were implemented in Perl, combined with shell scripting for automation. The genome

scan and downstream analyses were carried out in R (R Development Core Team 2008).

Results

Distribution of genome-wide Mutual Information (MI)

MI was calculated for five parapatric pair-wise comparisons and three comparisons of hither hierarchical population levels (two continent-wise comparisons and one global comparison). For the parapatric pair-wise comparisons, the genome-wide average MI within each population pair ranges from 0.058 to 0.126, in the ascending order of Us, G2, No, G1 and Ca. This ranking of the population pairs is consistent with the evaluation of genomic divergence by F_{ST} (Feulner et al. 2015). The distribution of genome-wide MI values within each comparison follows L-shaped distributions signifying that most loci are not differentiated (100 kb windows average, Figure 1a). Population pairs with lower average MI have more left-skewed distributions while population pairs with higher average MI have more widened distributions, which is also similar to the distributions of F_{ST} in these population pairs (Feulner et al. 2015).

The comparisons of higher hierarchical levels revealed lower genome-wide MI differentiation when contrasting ecotypes across a wider geographic scale, compared to the parapatric pair comparisons. The genome-wide average MI was 0.033 for the three European population pairs and 0.042 for the two North American population pairs (Figure 1b). When comparing ecotypes across all 5 population pairs together, an even lower genome-wide MI was obtained: 0.015. The shapes of MI distribution across genome-wide 100kb windows are also more left-skewed compared to European and North American comparisons (Figure 1c). The highest MI in the global comparison is 0.35 for SNPs and 0.048 for 100 kb windows.

For each of the eight comparisons (five parapatric pair-wise comparisons, two continent-wise comparisons and one comparison of all populations), outlier windows were identified as being significant from the permutation tests, as well

as the MI values at the top of empirical distribution. Parallel genetic changes accompanying ecotype differences across population pairs were investigated in two ways: by the overlap of outlier windows between pair-wise comparisons or by the outlier windows from higher hierarchical comparisons (continent-wise or global comparison). The former approach of overlapping outlier windows between pair-wise comparisons was also investigated using the F_{ST} approach (Feulner et al. 2015), making it possible to compare the results between MI and F_{ST}.

Outlier windows in pair-wise comparisons

To be comparable to the smaller window sizes used in the F_{ST} analyses, we also calculated MI averages for 10 kb windows, which results in 354 to 380 outlier windows for each population pair. We identified in a total of 88 shared outlier 10kb windows across population pairs, including 43 shared between the European population pairs (G1, G2 and No) and 12 shared between the North American population pairs (Ca and Us). Out of these 88 shared outlier windows identified by the MI 10kb approach, 24 out of the 47 shared outlier windows from F_{ST} approach (Table S5 in Feulner et al. 2015) were recovered. As half of the F_{ST} outlier windows were recovered, comparison between MI outlier windows and F_{ST} outlier windows indicate the general agreement between the two methods in identifying most extremely divergent regions shared by population pairs.

More than half of the 88 shared 10kb outlier windows were located consecutively on the chromosomes or clustered within 100kb between each other, suggestive of divergent genomic regions often exceeding the scale of 10 kb. Thus, we expanded the genome scan window size to 100kb. For each of the five parapatric pair-wise comparisons, we identified 19 or 20 100kb outlier windows. All outlier windows were unique to a population pair except for four outliers, which were shared within continents (Figure 2). Two outlier windows were shared by the two German population pairs (70^{th} window on group IV and 202^{nd} window on group IX) and two were shared by the two North American population pairs (195^{th} window on group IV and 214^{nd} window on group VII)

(Table 1). Except for the 202nd window on group IX shared by G1 and G2, other three of these four shared outlier windows (100kb) contain multiple outlier windows from the 10 kb approach, confirming that different window sizes give agreeing results.

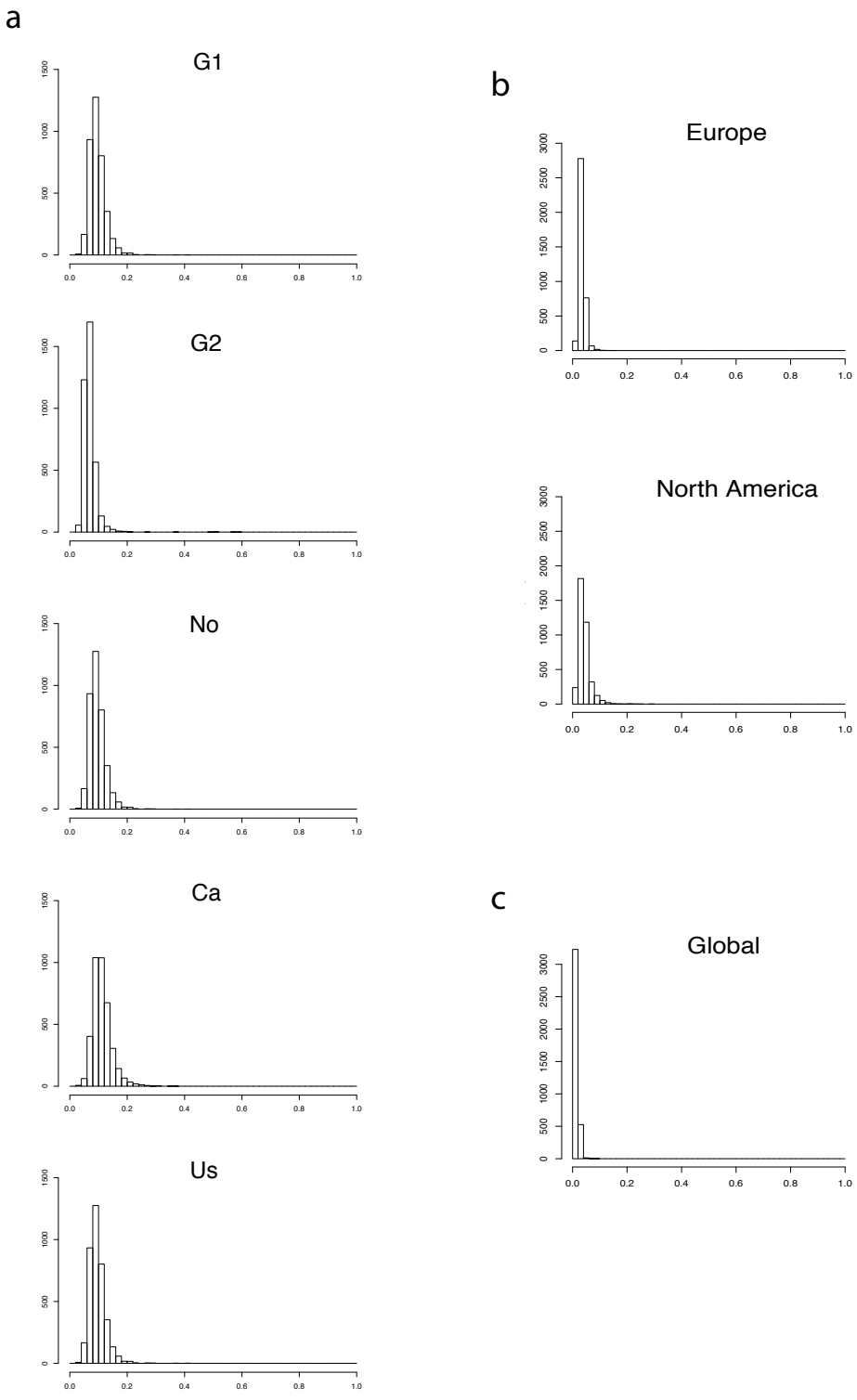


Figure 1: Distribution of MI (100 kb windows average) in (a) 5 parapatric pair-wise comparisons, (b) in continent-wise comparisons, and (c) in all-population comparison.

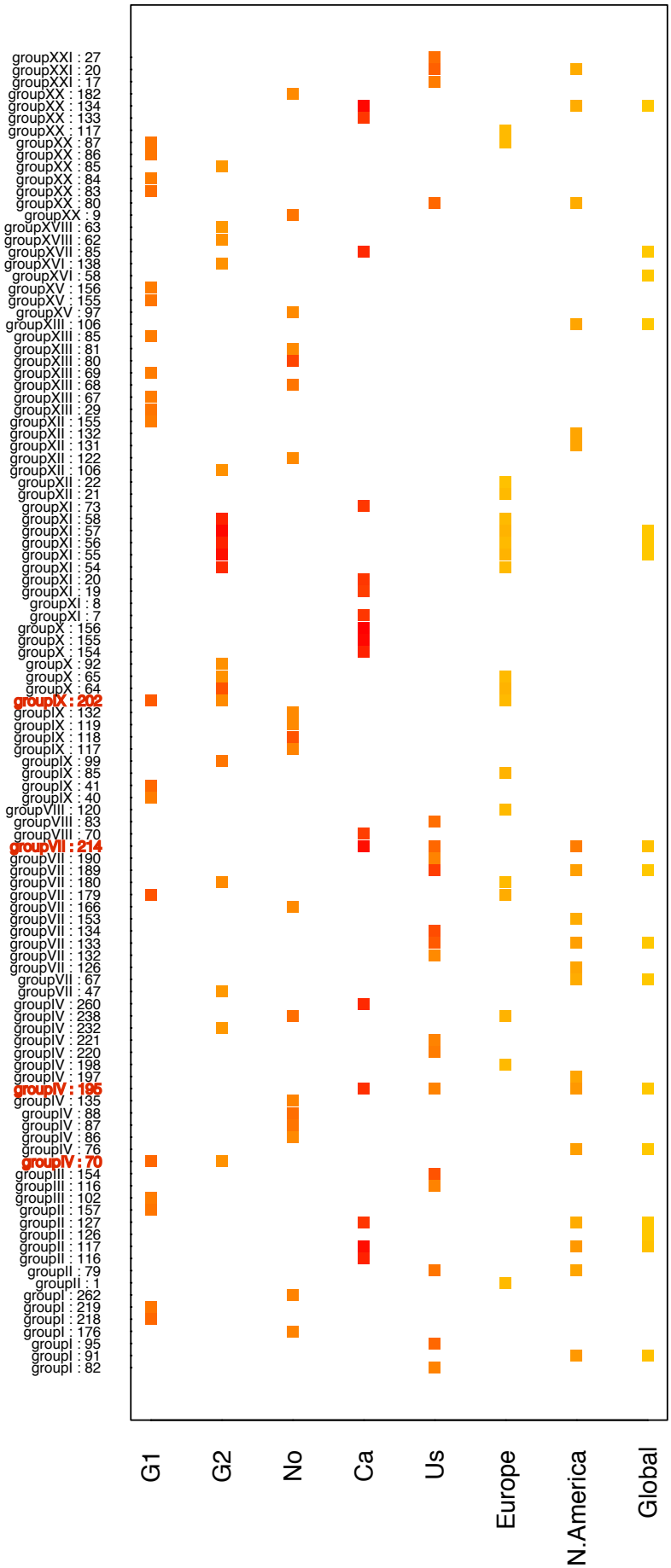

groupXXI : 27
groupXXI : 20
groupXXI : 17
groupXX : 182
groupXX : 134
groupXX : 133
groupXX : 117
groupXX : 87
groupXX : 86
groupXX : 85
groupXX : 84
groupXX : 83
groupXX : 80
groupXX : 9
groupXVIII : 63
groupXVIII : 62
groupXVII : 85
groupXVI : 138
groupXVI : 58
groupXV : 156
groupXV : 155
groupXV : 97
groupXIII : 106
groupXIII : 85
groupXIII : 81
groupXIII : 80
groupXIII : 69
groupXIII : 68
groupXIII : 67
groupXIII : 29
groupXII : 155
groupXII : 132
groupXII : 131
groupXII : 122
groupXII : 106
groupXII : 22
groupXII : 21
groupXI : 73
groupXI : 58
groupXI : 57
groupXI : 56
groupXI : 55
groupXI : 54
groupXI : 20
groupXI : 19
groupXI : 8
groupXI : 7
groupX : 156
groupX : 155
groupX : 154
groupX : 92
groupX : 65
groupX : 64
groupX : 202
groupIX : 132
groupIX : 119
groupIX : 118
groupIX : 117
groupIX : 99
groupIX : 85
groupIX : 41
groupIX : 40
groupVIII : 120
groupVIII : 83
groupVIII : 70
groupVII : 214
groupVII : 190
groupVII : 189
groupVII : 180
groupVII : 179
groupVII : 166
groupVII : 153
groupVII : 134
groupVII : 133
groupVII : 132
groupVII : 126
groupVII : 67
groupVII : 47
groupIV : 260
groupIV : 238
groupIV : 232
groupIV : 221
groupIV : 220
groupIV : 198
groupIV : 197
groupIV : 195
groupIV : 135
groupIV : 88
groupIV : 87
groupIV : 86
groupIV : 76
groupIV : 70
groupIII : 154
groupIII : 116
groupIII : 102
groupII : 157
groupII : 127
groupII : 126
groupII : 117
groupII : 116
groupII : 79
groupII : 1
groupI : 262
groupI : 219
groupI : 218
groupI : 176
groupI : 95
groupI : 91
groupI : 82
G1
G2
No
Ca
Us
Europe
N.America
Global

Figure 2: Summary of outlier windows for (a) pair-wise, (b) continent-wise and (c) all-population comparisons. The joint set of outlier windows from all comparisons are listed on the left, with the ordinal number of the 100 kb window indicated with the linkage group number. Red labels indicate shared outlier windows by at least two parapatric pair-wise comparisons. Only outlier windows in the corresponding comparisons are color-coded. The color code represents the average MI values from 0 (light yellow) to maximal 0.64

Table 1 Overlapped outlier windows and the annotated genes and GO terms.

Chromosome	Position (100 kb)	Comparisons (MI values)	Genes	GO terms
group IV	70	G1 (0.304), G2 (0.181), Euro (0.088)	-	-
group IV	195	Ca (0.471), Us (0.235), N.Am (0.196), All (0.048)	LRRN3 (leucine rich repeat neuronal 3b) ENSGACG00000018927	
group VII	214	Ca (0.575), Us (0.312), N.Am (0.290), All (0.062)	slc47a2 (2 of 2) (solute carrier family 47 (multidrug and toxin extrusion), member 1) slc47a1 (1 of 3) slc47a1 (2 of 3) slc47a1 (3 of 3) TMIGD1	Transmembrane transport, drug transmembrane transport, membrane, drug transmembrane transport activity, antiportor

			(transmembrane and immunoglobulin domain containing 1)	activity
group IX	202	G1 (0.354), G2 (0.211), Euro (0.086)	ENSGACG00000018927 SLC29A1 (2 of 2) solute carrier family 29 (equilibrative nucleoside transporter), member 1b	

Table 2: GO enrichment results among genes in overlapped outlier windows.

GO ID	GO terms	Annotated	Significant	Expected	P value	Adjusted-p
biological process						
GO:0006855	drug transmembrane transport	7	4	0	5.2e-13	1.2 e-09
GO:0015893	drug transport	8	4	0	1.0e-12	1.2 e-09
GO:0042493	response to drug	8	4	0	1.0e-12	1.2 e-09
GO:0042221	response to chemical stimulus	240	4	0.13	2.0e-06	0.00165
GO:0055085	transmembrane transport	594	4	0.31	7.3e-05	0.049
molecular function						

GO:0015238	drug transmembrane transporter activity	8	4	0	2.0e-12	2.5e-09
GO:0015297	antiporter activity	41	4	0.02	2.8e-09	1.8e-06
GO:0015291	secondary active transmembrane transport	158	4	0.08	6.8e-07	0.00029
GO:0022804	active transmembrane transporter activity	260	4	0.14	5.0e-06	0.0016
GO:0022857	transmembrane transporter activity	775	5	0.41	1.3e-05	0.0034
GO:0005215	transporter activity	919	5	0.49	3.0e-05	0.0066

Genes in the shared outlier windows can provide functional information about the advantageous alleles repeatedly involved in adaptation to lake or river environments. Genes in the four aforementioned shared 100kb outlier windows by two German or by two North American population pairs were overrepresented with functions involved in transmembrane transport (Table 2). These four windows contain a total of 7 genes encoding transmembrane proteins, except for the 70th 100kb window on group IV, which does not contain any annotated genes at all. The 195th window on group IV contains a leucine rich repeat neuronal 3 gene (LRRN3, ENSGACG00000018926), which has an immunoglobulin domain and a transmembrane region. The 202th window on group IX contains a solute carrier family 29 gene (slc29a1, ENSGACG00000020021), which is likely an equilibrative nucleoside transporter

(Baldwin et al. 2004). The 214th window on group VII contain 4 solute carrier family 47 genes (slc47a1 or slc47a2) that are annotated as multidrug and toxin extrusion proteins. This window also contains a gene with transmembrane and immunoglobulin domains (tmigd1, ENSGACG00000020616).

Outlier windows in continent-wise and global comparisons

The four aforementioned windows shared by pair-wise comparisons were also identified as outlier windows within European and North American continent-wise comparisons, respectively (Table 1). The continent-wise comparisons also identified other outlier windows, most of which were identified as outlier windows only in one pair-wise comparison, showing that high MI in one population pair can dominate the signal in the continent (Figure 2). The European comparison and the North American comparison have no overlap in outlier windows. When comparing all populations together, 17 out of the 19 windows with highest MI (within 0.5% top of empirical distribution) were found significant from the permutation test and identified as outliers. The MI values of these outlier windows range from 0.029 to 0.048. Nevertheless, 3 of the 17 outlier windows were also identified in the European comparison and another 11 in the North American comparison. The two outlier windows shared between Ca and Us were outlier windows in the North American comparison and in the global comparison. One of these outliers (214th window on group VII) has the highest average MI (0.048) from the global comparison, suggesting the two North American population pairs drive the global signal. Two outlier windows in the global comparison were not found as outliers in any other comparison (pair-wise or continent-wise). One is the 126th window on group II, which has MI values from 0.076 (Us), 0.144 (G2), 0.155 (G1), 0.160 (No) to 0.236 (Ca) in the parapatirc pair-wise comparisons. These values of MI are not outliers in the pair-wise comparisons, but all together contribute to make an outlier MI of 0.029 in the global comparison. Similarly, the 58th window on group XVI is MI values from 0.047 (G2), 0.070 (Us), 0.079 (No), 0.157 (G1) to 0.300 (Ca), but has a global MI of 0.030 as an outlier window.

Discussion

Using mutual information (MI) as an evaluation of genetic sequence divergence between lake and river stickleback ecotypes, we investigated parallel genomic divergence across replicated population pairs. The parapatric pair-wise comparisons revealed four outlier windows shared between pairs, which were also identified as outlier windows in their respective continent-wise comparisons. No outlier windows were shared between the two continent-wise comparisons, and the global comparison across continents revealed 17 outlier windows but at most with MI of 0.048. Taken together, the comparisons of lake and river stickleback ecotypes on different geographical scales (parapatric pair-wise, continent-wise and global) suggest some genetic parallelism within continents but little on a global scale.

The parapatric pair-wise comparisons revealed non differentiation in the majority of genomic regions between ecotypes, which is in line with the knowledge that the parapatric population pairs have newly diverged after the last glaciation, and that the possible continuous gene flow between the adjacent populations could keep homogenizing the genomic background. The few outlier genomic regions that show the most differentiation between adjacent populations, i.e. high MI, are potentially involved in adaptation to the lake or/and river habitat environments. The lake and river habitats are distinct environments both in abiotic and biotic aspects, differing in water flow regime, temperature, light sheltering, food resources, predator presence and parasite communities (Kalbe et al. 2002; Eizaguirre et al. 2011). The environmental differences between lake and river habitats may act as divergent selective agents on the standing genetic variation in the ancestral sticklebacks, driving the frequency of adaptive genotypes to change and differentiate between population dwelling different habitat environments (Barrett & Schluter 2008; Eizagiurre et al, 2009; Feulner et al. 2013). Standing genetic variants pre-existing in the ancestral sticklebacks could be readily available when adapting to new environments. Especially if such variants exist in high frequency, it would increase the probability of gene reuse and parallel evolution compared to new mutations (Schluter et al. 2004).

Taking advantage of the wide distribution and adaptive radiation of three-spined sticklebacks, we investigated whether lake and river ecotypes in independent population pairs had same genomic regions being repeatedly selected. Pair-wise comparisons of allele frequency divergence as determined by MI revealed that most outlier windows were population specific, which is consistent with the F_{ST} approach applied on the same dataset (Feulner et al. 2015). Four outlier windows were found shared between population pairs within the same continent, which overlapped with the shared F_{ST} outlier windows, suggesting robustness of both methods for the parapatric pair-wise comparisons. One of the shared outlier windows from the two North American population pairs (group VII, 214th 100kb window) also has the highest MI value (0.048) from the global comparison. This window predominantly contains transmembrane protein genes, which might serve as a responding mechanism to the environments. The other regions that were shared across population pairs within a continent also contain transmembrane protein genes, in addition to some immune function genes, suggesting functions in defense against pathogens and parasites in the environments.

Besides overlapping outlier windows from pair-wise comparisons, another way to evaluate genomic regions with parallel evolution is to evaluate lake-river divergence at higher hierarchical levels: continent-wise or global-wise. The advantages of comparing multiple pairs of lake and river ecotypes together compared to overlapping outlier windows from pair-wise comparisons include that it looks for the exact same variants (SNPs) repeatedly diverged across population pairs instead of averaging within windows, taking into account the direction of allele frequency changes. As F_{ST} is based on interbreeding coefficients at different levels of hierarchically subdivided populations (Slatkin & Voelm 1990), our goal to compare lake and river ecotypes across multiple population pairs would violate the population structure and lead to inappropriate estimation of interbreeding coefficients. Instead, the concept of MI from Shannon's information theory, which measures the concordance of genotypes with ecotypes, is free of such theoretical assumptions

(Donaldon-Matasci et al. 2010). Our MI analyses of continent-wise and global comparisons rediscovered the four outlier windows shared by pair-wise comparisons, further suggesting that not only do the ecotypes diverge at these regions, but they do so using the same alleles. This suggests that particular alleles are beneficial and differentially selected in these different habitats. The lack of overlap of outlier windows between the continent-wise comparisons of Europe and North America, and the prominent low MI values in the global comparison suggest lack of parallelism across continents both at the SNP level and at the region level.

The lack of parallel genetic evolution across continents can be explained by the different origins of the freshwater populations. The European and North American freshwater stickleback populations that were used in this study are derived from the Atlantic and Pacific marine ancestral populations, respectively. Thus, the divergence between lake and river stickleback ecotypes in different continents started from different pools of standing genetic variation, and the genetic variation under selection had different genomic backgrounds to interact with, making the reuse of the same genes across continents unlikely. In contrast, repeated gene usage is more likely within population pairs that share standing genetic variation from recent ancestors. These speculations are consistent with previous estimations that the probability of gene reuse is higher when populations or species are more closely related and share the same ancestors (Conte et al. 2012). While studies contrasting marine and freshwater sticklebacks (Jones et al. 2012; Hohenlohe et al. 2010) found plenty of parallel genetic changes, no highly differentiated regions were shared between continents in cases contrasting lake-river divergence (this study, Feulner et al. 2015, Deagle et al. 2011; Roesti et al. 2012b).

Taken together, in this chapter we evaluated parallel genetic divergence between stickleback ecotypes using a novel approach based on MI from information theory, which was not only employed for parapatric pair-wise comparisons, but also enabled population comparisons at higher hierarchical levels: across populations within continents and across continents. The pair-wise comparisons

confirmed previous results based on classical population genetic measurements (F_{ST}). These comparisons at different population levels together revealed a few putative parallel genomic regions but little parallelism across continents. Interpreting these few parallel genomic regions as our best candidates to understand the targeted functions of parallel adaptation, functional annotations such as transmembrane protein genes and immune functions suggest response mechanisms to the environments playing an important role in divergence between lake and river sticklebacks.

Chapter II Transcriptome profiling of immune tissues reveals habitat-specific gene expression between lake and river sticklebacks

Yun Huang*[1], Frédéric JJ Chain[1,2], Mahesh Panchal[1,3,4], Christophe Eizaguirre[5], Martin Kalbe[1], Tobias L Lenz[1] , Irene E Samonte[1], Monika Stoll[6], Erich Bornberg-Bauer[7], Thorsten BH Reusch[8], Manfred Milinski[1], Philine GD Feulner[1,9]

[1] Department of Evolutionary Ecology, Max Planck Institute for Evolutionary Biology, 24306 Plön, Germany

[2] Department of Biology, McGill University, Montreal, QC, H3A 1B1, Canada

[3] Bioinformatics Infrastructures for Life Sciences (BILS), Uppsala Biomedicinska Centrum (BMC), Husargatan 3, 751 23 Uppsala, Sweden

[4] Institute of Medical Biochemistry and Microbiology, Uppsala Biomedicinska Centrum (BMC), Husargatan 3, 751 23 Uppsala, Sweden

[5]School of Biological and Chemical Sciences, Queen Mary University of London, E1 4NS London, UK.

[6]Institute of Human Genetics, Genetic Epidemiology, Westfälische Wilhelms University, 48149 Münster, Germany

[7] Institute for Evolution and Biodiversity, Evolutionary Bioinformatics, Westfälische Wilhelms University, 48149 Münster, Germany

[8]Evolutionary Ecology of Marine Fishes, GEOMAR Helmholtz Centre for Ocean Research Kiel, 24105 Kiel, Germany

[9]Department of Fish Ecology and Evolution, Eawag, Swiss Federal Institute of Aquatic Science and Technology, Center for Ecology, Evolution and Biogeochemistry, 6047 Kastanienbaum, Switzerland

*Correspondence to: yun.huang@evolbio.mpg.de

This chapter has been published in *Molecular Ecology* (2016) 25:943-958.

Abstract

The observation of habitat-specific phenotypes suggests the action of natural selection. The three-spined stickleback (*Gasterosteus aculeatus*) has repeatedly colonized and adapted to diverse freshwater habitats across the northern hemisphere since the last glaciation, while giving rise to recurring phenotypes associated with specific habitats. Parapatric lake and river populations of sticklebacks harbour distinct parasite communities, a factor proposed to contribute to adaptive differentiation between these ecotypes. However, little is known about the transcriptional response to the distinct parasite pressure of those fish in a natural setting. Here, we sampled wild-caught sticklebacks across four geographical locations from lake and river habitats differing in their parasite load. We compared gene expression profiles between lake and river populations using 77 whole-transcriptome libraries from two immune-relevant tissues, the head kidney and the spleen. Differential expression analyses revealed 139 genes with habitat-specific expression patterns across the sampled population pairs. Amongst the 139 differentially expressed genes, 8 are annotated with an immune function and 42 have been identified as differentially expressed in previous experimental studies in which fish have been immune challenged. Together these findings reinforce the hypothesis that parasites contribute to adaptation of sticklebacks in lake and river habitats.

Introduction

The repeated occurrence of similar phenotypes associated with a distinct habitat is often attributed to the direct effect of natural selection (Elmer & Meyer 2011). Parallel phenotypic evolution among populations from geographically distant but ecologically similar habitats, referred to here as habitat-specific phenotypes, are thought to reflect the advantages of those phenotypes in their respective habitat (Savolainen et al. 2013). Numerous examples have been documented including pharyngeal jaw and thick lips in cichlids (Albertson et al. 2005; Colombo et al. 2013), similar ecotype morphs of anolis lizards (Harmon et al. 2005; Losos et al. 1998), habitat-specific pigmentation in isopods (Hargeby et al. 2004), repeated ecotypes with distinct shell sizes in the periwinkle snail (Butlin et al. 2014), and repeated differences of body depth and gill raker numbers between lake and stream sticklebacks (Berner et al. 2008; Kaeuffer et al. 2012; Lucek et al. 2014). Although phenotypic plasticity can contribute to such habitat-specific phenotypes (Machado-Schiaffino et al. 2014; Moser et al. 2015; Muschick et al. 2012), some of these traits have been shown to be genetically determined and under adaptive evolution (Albertson et al. 2005; Colombo et al. 2013; Hargeby et al. 2004). Adaptive genetic changes include those that result from polymorphisms that alter protein structures (Ffrench-Constant et al. 1993; Hoekstra et al. 2006; Protas et al. 2006) as well as those that influence phenotypes via regulation of gene expression (Chan et al. 2010; Rebeiz et al. 2009). Gene expression has been associated with adaptive changes in morphological and physiological changes (Harrison et al. 2012; Manceau et al. 2011; Rebeiz et al. 2009) and is believed to contribute to adaptive divergence in natural populations (Pavey et al. 2010).

As gene expression bridges the underlying genotype to the ultimate morphological and physiological phenotypes, it can be considered as an extended molecular phenotype (Ranz & Machado 2006). Hence, it is interesting to evaluate whether or not gene expression patterns differ between contrasting habitats and if so whether they hold across geographically distant populations. Such habitat-specific gene expression could arise due to several factors, such as

genetically determined expression patterns among similar habitat types (ecotypes), as well plastic responses to extrinsic environmental conditions specific to a habitat. Aside from other mechanisms that might control regulation of transcription such as epigenetics, genetic studies have demonstrated variable degrees of heritability of gene expression and have for some phenotypes revealed the genetic basis underlying expression differences (Gibson & Weir 2005; Gilad et al. 2008; Stamatoyannopoulos 2004). There are examples of mutations affecting cis- and trans-regulatory regions in the genome that silence or dramatically shift gene expression, including single nucleotide polymorphisms (SNPs) (Cheung & Spielman 2009; Fraser 2013), copy number variations (CNVs) (Haraksingh & Snyder 2013) and tandem repeats (Gemayel et al. 2010). Genomic changes in regulatory regions can alter the efficiency of transcription factors and thus affect expression of adjacent or remote genes. In sticklebacks for example, frequent independent deletion events in the enhancer of Pitx1 suppress expression of the gene and result in repeated pelvic reduction in freshwater populations (Chan et al. 2010). Besides its heritable (genetic) component, gene expression is also a versatile phenotype that dynamically responds to changes in the environment (Gibson 2008) and holds the potential to facilitate plasticity to buffer against environmental changes (Franssen et al. 2011; Morris et al. 2014; Whitehead 2012). Despite the variability introduced by uncontrollable environmental factors, studies of gene expression in wild-caught populations offer the opportunity to estimate the physiological responses of organisms in their environment, potentially providing insight into the role of gene expression variation in adaptation and acclimation to environmental stresses through genetic or plastic changes (Cheviron et al. 2008).

The repeated and independent postglacial colonization history of the three-spined stickleback (*Gasterosteus aculeatus*) makes it a powerful study system to investigate habitat-specific phenotypic evolution. Sticklebacks inhabit various marine and freshwater habitats across the northern hemisphere (MacKinnon & Rundle 2002), a distribution likely attributable to rapid adaptation from extensive standing genetic variation (Barrett & Schluter 2008; Eizaguirre et al. 2012a). Genetically diverged but geographically adjacent lake

and river population pairs exhibit consistent morphological differentiation across multiple pairs, such as divergence for body depth and gill raker number (Berner et al. 2008; Kaeuffer et al. 2012; Lucek et al. 2014). These lake and river populations are also often referred to as ecotypes (Reusch et al. 2001). Many ecological factors differ between lake and river habitats, such as flow regime, temperature, food resource and predator communities, all contributing to the differentiation of lake and river stickleback ecotypes, e.g. in foraging traits (Berner et al. 2010) and anti-predator traits (Lucek et al. 2014). Another important ecological difference between lakes and rivers is the locally distinct parasite communities (Eizaguirre et al. 2011; Kalbe et al. 2002; Karnonen et al. 2015). Besides harbouring different species of parasites between ecotypes, lake fish commonly have a higher parasite load than river fish comparing parapatric population pairs (Eizaguirre et al. 2011), and higher immuno-competence (Scharsack et al. 2007). Lake fish also exhibit a higher diversity in the major histocompatibility complex (MHC) (Eizaguirre et al. 2011), believed to be a result of local adaptation (Eizaguirre et al. 2012b; Eizaguirre et al. 2009). Distinct immune expression patterns between lake and river individuals were detected upon multiple experimental parasite exposure of laboratory-bred sticklebacks (Lenz et al. 2013). Altogether, these studies suggest that parasites play an important role in the differentiation of lake and river ecotypes by shaping the diversity and expression patterns of immune-related genes. It is, however, not yet known whether the generality of these patterns holds in multiple lake-river systems under natural conditions.

In this study we performed an extensive transcriptomic survey using an RNAseq approach across four parapatric lake and river stickleback population pairs to investigate patterns of habitat-specific gene expression. We used two major organs involved in immune response, the head kidney and the spleen. Differential expression analysis was performed between fish from lake and river habitats, and results were compared to the differentially expressed genes between laboratory-bred individuals in controlled parasite infection experiments (Haase et al. 2014; Lenz et al. 2013). Our study describes gene

expression differences in an ecological framework, highlighting habitat-specific expression of genes that might be involved in adaptation.

Materials and Methods

Sampling

Three-spined sticklebacks were sampled in 2010 for genomic studies (Chain et al. 2014; Feulner et al. 2015), from which four parapatric lake-river population pairs were used in this study. These included two independent drainages from Germany: Großer Plöner See lake (G1_L) and Malenter Au river (G1_R), Westensee lake (G2_L) and Eider river (G2_R), one pair from Norway: Skogseidvatn lake (No_L) and Orraelva river (No_R), and one pair from Canada: Misty Lake (Ca_L) and Misty Stream (Ca_R) (See Table 1). All these lake-river population pairs are significantly differentiated from each other, with a mean genome-wide FST ranging between 0.11 and 0.28 (for more detailed information about sampling sites and genetic differentiation between the populations, see Feulner et al. 2015). The two population pairs from Germany were sampled in May while the Norwegian and Canadian populations were sampled in September. About 20 individual fish per site were caught using dip nets or minnow traps and kept alive for a few hours in the water from where they were sampled until being euthanized using MS222 and dissection. For each population pair, the fish were treated identically after capture and lake fish and river fish were alternately dissected. Fish standard length and weight were recorded and macroparasites screened following established procedures for three-spined sticklebacks (Kalbe et al. 2002) (Supplementary Table 1). Immediately after euthanasia, the whole head kidneys and spleens were dissected out and preserved in RNAlater (Sigma-Aldrich) for later transcriptomic library preparation. These are the main immune organs in teleost fish and are commonly used for immunological studies (Press & Evensen 1999). Six individuals (3 males and 3 females, except No_L with 4 males and 2 females) were selected for transcriptomic sequencing per sampling site. Fish selection was performed ignoring parasite screening results, but was non random to ensure an equal sex distribution for each population and

with a preference for larger fish to guaranty sufficient yield of RNA. Body weights of the selected fish suggest that all fish were older than 1 year (Supplementary Table 1).

RNA library preparation and sequencing

Total RNA (using the entire tissue dissected) was extracted from preserved samples using NucleoSpin® RNA (Mackerey-Nagel) and reverse transcribed to cDNA using Omniscript RT kits (Qiagen). RNA was quantified with NanoDrop and Bioanalyzer and ~1µg of RNA in a concentration of 20ng/µL was used for library construction. A few samples with poor RNA quality were excluded before constructing 77 libraries. Therefore, sample sizes per population vary between 3 and 6 individuals (Table 1). TruSeq RNA sample preparation kit (Illumina) was used for paired-end library construction according to the manufacturer's instructions. Each sample was barcoded with a unique sequence index tag and pools of 12 different barcoded samples were loaded in 8 lanes of a single flow-cell of Illumina HiScanSQ machine.

Read filtering and mapping

Raw reads were quality filtered before read mapping in the following steps. All raw reads output to fastq files were 101 base pairs (bp) in length. Sequencing adaptors were removed using SeqPrep 0.4 (https://github.com/jstjohn/SeqPrep). PrinSeq 0.20.3 lite (Schmieder & Edwards 2011) was used to trim the read tails with a PHRED quality score below 20 as well as poly-A tails longer than 10 bp. We kept read-pairs for which both reads were longer than 60 bp after trimming. After filtering, read lengths varied from 60 to 101 bp, with about 60% of the reads exhibiting the initial 101 bp length. Exact duplicates of both paired-ends were removed with PrinSeq. The remaining quality-filtered reads were aligned against the stickleback reference genome from Ensembl version 68 (Flicek et al. 2012) using Tophat2 v2.0.13 (Kim et al. 2013) with default settings. HTSeq 0.5.4p5 (Anders et al. 2014) was used to quantify read count for each gene using Ensembl gene annotations (version 68) using default settings except for excluding reads with alignment quality below 5.

Gene expression analyses

Gene expression across all samples was evaluated with the Bioconductor package EdgeR 3.4.2 (Robinson et al. 2010). First, weakly expressed genes were filtered out when they had less than 1 read per million in half (38) of the 77 samples (Anders et al. 2013). All libraries were then simultaneously normalized with the trimmed mean of M-value (TMM) method (Robinson & Oshlack 2010), implemented in the EdgeR package. The TMM method computes the scaling factors as the weighted mean of log fold changes for the majority of genes between libraries, based on the assumption that the majority of genes are not differentially expressed. After applying the TMM method most genes should have unified expression levels across individuals and the scaling factors for the libraries should be close to 1 (Dillies et al. 2012). Except for one head kidney library from G1_R with a scaling factor of 0.35, all other transcriptome libraries obtained scaling factors close to 1 (from 0.75 to 1.18, Supplementary Table 2). The outlier library had fewer genes expressed compared to other libraries (12769 versus 15735-17341). This indicates a distinct expression profile likely dominated by technical artifacts, and therefore this library was excluded from further analyses.

Next, the dispersion of the negative binomial distribution for the expression of each gene was estimated in EdgeR. It represents the biological coefficient of variation of a gene's expression. This was used to evaluate the expression variance where a high dispersion value indicates high variance of gene expression pattern among samples. A principal component analysis (PCA) was then performed in R 3.0.1 (R Development Core Team 2008) using prcomp function based on log-transformed normalized read counts of all 12222 expressed genes (across both tissues and after filtering out weakly expressed genes as mentioned above) to assess differences in gene expression across libraries (Figure 1).

To identify habitat-specific gene expression, i.e. the expression patterns that are similar within habitat types while significantly different between habitat types, we employed differential expression (DE) analyses that contrast lake and river

fish from all four population pairs. On the basis of the PCA result (Figure 1), DE analyses were performed separately for head kidney and spleen libraries in EdgeR. Because the PCA results suggest that the Canadian populations are substantially diverged from the European populations, the DE analyses were also performed only among the three European population pairs (those results are presented in the Supplement only). Hence, four DE analyses were performed (comparing gene expression in the head kidney across all four population pairs, in spleen across all four population pairs, in head kidney across only the three European population pairs, and in spleen across only the three European population pairs). Before conducting DE analyses, weakly expressed genes were filtered out to avoid bias in fold changes due to weak expression of some genes. Genes were filtered out from the DE analyses if they did not have at least 1 read per million in n of the samples, where n is the size of the smaller group (lake or river) in the DE comparisons (Anders et al. 2013). Libraries were re-normalized within each comparison group with the TMM method in EdgeR. A multi-factor design was used in a negative binomial generalized linear model, which accounts for the variation attributed to different population pairs as well as for the variation associated to the sex of the individuals (Expression~Habitat type + Population pair + Sex). The gene-wise dispersion was re-estimated based on the generalized linear model within each comparison group. For each tissue, the distribution of dispersion values were left-skewed with long tails, indicating that most genes had uniform expression, with a small proportion of genes having highly variable expression across individuals being compared (Supplementary Figure 1). We calculated the Pearson correlation of gene expression between all possible pairs of individuals within biological replicates (individuals of the same habitat, population pair, and sex) using count data in R. The overall average correlation of gene expression across all pairwise comparisons was 0.86 (first quartile: 0.81 and third quartile: 0.95). Likelihood ratio tests for the contrast coefficient (lake versus river) were performed and p-values were corrected for multiple testing using the Benjamini-Hochberg method (Benjamini & Hochberg 1995). Genes with corrected p-values smaller than 0.05 were categorized as differentially expressed genes (DE genes). In addition to performing all DE analyses in EdgeR as described above, DE analyses were also performed with the

default pipeline in the DESeq2 package 1.0.19 (Love et al. 2014) giving similar results (Supplementary Table 3).

Functional analyses

Out of 20.787 stickleback genes, 13568 are annotated with Gene Ontology (GO, (Ashburner et al. 2000)) terms in Ensembl version 80. We complemented this with 13.044 gene annotations acquired from the Zebrafish Model Organism Database (ZFIN, Howe et al. 2013) genes associated with stickleback Ensembl IDs, with annotation information from:

ftp://ftp.geneontology.org/pub/go/gene-associations/gene_association.zfin.gz

After merging all annotations, a total of 17081 out of 20787 stickleback genes were annotated with GO terms. We tested for the enrichment of GO terms in our DE gene sets with the Bioconductor package topGO (Alexa & Rahnenfuhrer 2010; Alexa et al. 2006), based on Fisher's exact tests. The gene pools against which we compared the DE gene sets were the genes having sufficient expression and entering the differential expression analyses (see gene expression analyses section above). Overrepresented GO terms were those with a multiple-test corrected p-value (Benjamini-Hochberg's false discovery rate, FDR) smaller than 0.05. To infer the potential involvement of the habitat-specific expressed genes in parasite defense in nature, we identified our DE genes that were also differential expressed in two previous laboratory-controlled parasite exposure experiments (Haase et al. 2014; Lenz et al. 2013).

Results

Qualitative description of expression patterns

For each of the 77 transcriptome libraries, an average of 6.5 million read pairs of 101 bp were produced. After adapter cleaning, quality trimming, and duplicate- and length-filtering, 92.78% of the reads remained for analyses (Supplementary Table 2). On average, 88.10% of the quality-filtered reads mapped to the reference genome and 2.71% of these mapped to multiple regions of the genome, which were subsequently excluded from further analyses. Out of a total of 22456

genes annotated in the stickleback genome (Ensembl version 68), an average of 16397 (+/-944) genes were found expressed. The median number of reads mapping back to each expressed gene was 60 read pairs (first quartile to third quartile: 13-166). The principal component analysis (PCA) clearly separated the two tissue types along the first principal component, which accounted for 41% of the variance observed in the dataset (Figure 1). Within the same tissue type, the second principal component (variance explained: 8%) clearly separated European samples from the Canadian samples.

Differential expression (DE) analyses

After filtering out weakly expressed genes (see Methods), 12105 genes expressed in head kidney and 12451 expressed in spleen were contrasted between lake and river ecotypes across all four population pairs. A total of 139 genes showed significant differential expression after correction for multiple testing (Figure 2). There were 73 DE genes in the head kidney, 74 DE genes in the spleen, and 8 of these genes were shared between both tissues (Supplementary Table 3). All 8 shared DE genes showed the same directional difference of expression between habitat types. A majority of the DE genes (75% in head kidney and 65% in spleen) showed higher expression in individuals from lakes than from rivers. Most of these same DE genes were identified using another commonly used software with default parameters (DESeq2: 70 out of 73 in the head kidney and 67 out of 74 in the spleen, Supplementary Table 3). Although the PCA analyses mentioned above suggested that the overall expression patterns of the European samples seemed distinct from the Canadian samples, a separate analysis of expression log fold changes between lake and river fish from the three European population pairs showed a strong positive correlation with that of all four population pairs together (linear regression, $R2=0.61$, $p<0.001$ for head kidney and $R2=0.82$, $p<0.001$ for spleen), and resulted in about half of the same DE genes (Supplementary Table 4). The 5 DE genes with the smallest adjusted p-value in the head kidney across all lake-river comparisons include 3 genes that have higher expression in lake fish (*leucine-rich repeat containing 17*, *ryanodine receptor 3*, and *colony-stimulating factor 1b*) and two that have higher expression in river fish (*cub and sushi*

multiple domains 3 and one uncharacterized protein coding gene ENSGACG00000000187). The 5 genes with smallest adjusted p-values in the spleen include three that have higher expression in lake fish (*solute carrier family 43, member 3b*, *actin binding LIM protein 1b*, and *complement factor D*) and two uncharacterized protein coding genes (ENSGACG00000000187 and ENSGACG00000012387) that have higher expression in river fish (see Supplementary Table 3 for all 139 DE genes identified).

Functional analyses of DE genes

GO annotations from Ensembl and the ZFIN database were available for 105 of the 139 DE genes (Supplementary Table 3). The DE genes in head kidney had no significant GO term enrichment, while the DE genes in spleen were enriched for collagen (GO:0005581, with 3 out of 18 genes annotated with this term in the gene pool), extracellular region (GO:0005576, with 8 out of 265 genes) and extracellular matrix part (GO:0044420, with 3 out of 20 genes). Applying a less stringent cut-off for DE genes (FDR<0.10) to test for enrichment of GO terms (FDR<0.05), only extracellular region (GO:0005576) remained significant in the spleen, with no additional terms found in both tissues. The top 50 GO terms from the enrichment analyses of original DE gene sets (FDR<0.05) are provided in the Dryad database (see Data Accessibility Section). To specifically investigate the differential expression of immune genes in the sampled immune-related tissues, a list of 1126 stickleback genes with putative immune functions was acquired from a previous study (Haase et al. 2014). Among the DE genes between lake and river fish, 3 of the 73 DE genes in the head kidney and 5 of the 74 DE genes in the spleen are putatively immune genes (Table 2). These included macrophage receptors, an interferon regulatory factor and a gene annotated with the functions of antigen processing and presentation and immune response.

While our analysis only detected very few immune function genes showing differential gene expression, the parasite survey of our sampled fish showed that lake fish harbor higher parasite loads than river fish (Supplementary Table 1). This has already been demonstrated previously using a larger sample size (Figure 1 in Feulner et al. 2015). To further investigate the role of parasite

infection and potential resistance in driving differential gene expression between lake and river habitats, we compared our results with two laboratory-controlled parasite exposure experiments that assessed gene expression in sticklebacks from the same German populations as used in our study. Lenz et al. (2013) described the transcriptional responses of laboratory-bred lake and river sticklebacks under either controlled or parasite-challenged conditions. That study used three parasites that are found in the natural environment of those fish: *Diplostomum pseudospathaceum*, *Anguillicola crassus*, and *Camallanus lacustris*. These parasites were also found in our sampled fish (see discussion and Supplementary Table 1). Out of 166 DE genes between twice parasite-exposed lake and river fish (Lenz et al. 2013), 51 and 73 genes showed the same directional differences of expression between habitat types in our study among all lake-river population pairs, in the head kidney and in the spleen respectively. Some of the differences between the two studies are likely due to that the majority of DE genes in Lenz et al. 2013 were highly expressed in river fish as they are exposed to equal dosage of parasites compared to lake fish, while in our study the majority of DE genes were highly expressed in lake fish as the river fish were exposed to less parasites in nature. Nevertheless, amongst those genes with same directional differences, one gene *methyltransferase like 13* (*mettl13*) was also identified significantly differentially expressed in our study (Table 3, also see Discussion for more details). In addition, 10 of the 1057 DE genes between control and parasite-challenged fish (Lenz et al. 2013) overlapped with our set of DE genes (4 in the head kidney and 6 in the spleen). In another recent parasite infection study, laboratory-bred lake sticklebacks (from the G1_L population) were challenged with the trematode *Diplostomum pseudospathaceum* (Haase et al. 2014), and DE was assessed in the head kidney and in the gill. Out of 1060 DE genes between control and challenged fish in the head kidney (Haase et al. 2014), 6 overlapped with the DE genes from our study (all in the spleen). Out of 1415 DE genes in the gill (Haase et al. 2014), 25 overlapped with our set of DE genes (12 in the head kidney and 14 in the spleen, including 1 in both tissues, Table 3).

Discussion

Habitat-specific expression

This study investigated transcriptional profiles of three-spined sticklebacks from contrasting lake and river habitats across a wide geographical scale. Physical and ecological differences between lake and river habitats, consisting of differences in flow regime, vegetation, food resources, and parasite communities among others, can influence individual fitness, behaviour, life history, morphology and physiology. Studies contrasting lake and river sticklebacks have mainly focused on their morphology (Berner et al. 2010; Lucek et al. 2014) and genomic variation (Chain et al. 2014; Deagle et al. 2012; Feulner et al. 2015; Roesti et al. 2012). Here, we evaluated how lake and river ecotypes differ in gene expression profiles in their natural environments. We have identified habitat-specific gene expression patterns, i.e. differential expression between habitats across four lake-river pairs, three from European locations and one from Canada. For differentially expressed genes, fish from the same habitat have a similar expression, which is distinct from the expression in fish from the contrasting habitat. These habitat-specific expression patterns suggest that a part of the transcriptome (about 1%) is shaped by the global environmental contrast across all lake-river pairs, although a larger fraction may be affected by local habitat differences within a given population pair or expressed in other tissues or during a different season or ontogenetic stage. These findings add to the growing discussion of parallelism at the regulatory level between contrasting ecotypes and morphs (Derome et al. 2006; Manousaki et al. 2013; Pavey et al. 2011).

Plasticity and heritability of gene expression

A combination of evolutionary mechanisms could be shaping the habitat-specific expression patterns observed in this study. Freshwater sticklebacks likely possess the innate ability to regulate certain genes in acclimating to the different conditions in lakes and rivers (Stutz et al. 2015). This plasticity could result in habitat-specific expression patterns. Alternatively, differential expression across habitats might also reveal adaptive genetic differences between lake and river fish. These alternative explanations for habitat-specific patterns are by no means

mutually exclusive, and may both contribute to shape the gene expression profiles of lake and river sticklebacks. Setting our study into the context of previous findings, we further evaluated these explanations. Using the same individuals from this study (as well as additional individuals), recent genomic studies have shown little evidence for sequence-based habitat-specific patterns using genome scan approaches with single nucleotide polymorphisms (SNPs; Feulner et al. 2015) and with copy number variations (Chain et al. 2014). Hence, from a genomic perspective, despite significant differentiation between lake and river sticklebacks at a regional scale and across a wider continental scale (Deagle et al. 2012; Feulner et al. 2015; Roesti et al. 2012), there is little evidence for parallel genetic differentiation between lake and river sticklebacks across the distribution area of the fish. In other words, genetic differences between freshwater ecotypes of sticklebacks are for the large part not shared across population pairs, whereas here we identified several genes with habitat-specific gene expression. This discrepancy is consistent with the observation that phenotypes are similar amongst lake-river populations while the genetic basis is different (Deagle et al. 2012; Feulner et al. 2015; Kaeuffer et al. 2012). Gene expression, which bridges the underlying genetic basis and the ultimate phenotypes, might contribute to the understanding of the discrepancy between phenotypes and genotypes. Habitat-specific expression patterns could be controlled by various trans-regulatory elements from different genomic sources in different populations. Another explanation is that pathways regulating expression might be triggered at different steps in signaling cascades and therefore leave distinct signatures in the genomes of different populations (Pritchard et al. 2010). Based on controlled laboratory studies, there is evidence that expression differences in sticklebacks can be largely heritable (Leder et al. 2014). In addition, a laboratory-controlled experiment in which laboratory-bred G1_L and G1_R sticklebacks exhibited different transcriptional responses to parasite exposure suggested that the genetic background plays an important role in differential gene expression between fish ecotypes (Lenz et al. 2013). It is interesting that this differentiation between lake and river fish was most pronounced in their adaptive immune response (triggered upon 2nd exposure) to parasites, most likely resembling the differences we are observing in nature,

where the fish are very likely to have multiple encounters with parasites. In light of these studies, adaptive genetic differences between lake and river sticklebacks appear to be a likely explanation for habitat-specific expression patterns. However, a reciprocal transplant experiment suggested that environmentally induced plasticity strongly affects the expression of some carefully selected immune genes (Stutz et al. 2015). Hence, plasticity in gene expression might have also shaped the habitat-specific expression pattern of some of the genes identified in this study.

Immunological relevance of DE genes

Large-scale observational studies such as the current one are complementary to experimental studies in general, and here to the stickleback system in particular. Previous studies on sticklebacks in German lake-river systems highlighted that lake fish harbour higher parasite loads than river fish in terms of intensity and species diversity (Eizaguirre et al. 2012b; Eizaguirre et al. 2011; Kalbe et al. 2002). This trend of contrasting parasite loads was further confirmed across a wide geographic range including all populations used in our study (Feulner et al. 2015). Experiments have established that lake and river sticklebacks have differences in immune-competence due to habitat-specific adaptation to the distinct parasite communities (Scharsack et al. 2007). It was further investigated that genetic differences in MHC genotypes between lake and river fish provide a basis for parasite-mediated local adaptation (Eizaguirre et al. 2012a; Eizaguirre et al. 2011) following the idea that parasite resistance could represent a magic trait involved in speciation (Eizaguirre et al. 2009). As the differences in parasite pressure between niches could be a force driving divergent adaptation in lake and river sticklebacks, we surveyed gene expression in immune tissues with a specific focus on genes involved in immune functions. Across the 139 candidate genes, we found 3 putative immune genes in the head kidney and 5 in the spleen with habitat-specific expression patterns (Table 2). We found that genes with an immune function were not overrepresented, which indicates that under natural conditions, other factors besides parasites and immunity also contribute to the differentiation between ecotypes. The overrepresented GO terms from these habitat-specific expressed genes suggest the gene products are often

extracellular components, such as collagen-structured proteins. Given the generic GO terms, their contribution to habitat-specific adaptation is open to speculation. Nevertheless, a detailed examination of the DE genes showing most significant expression differences (with smallest adjusted p-values) between lakes and rivers revealed some associations with immune-related functions. One of the genes that is highly expressed in lake fish and differentially expressed in both the head kidney and in the spleen is *colony-stimulating factor 1b* (*csf1b*), which is involved in macrophage production and differentiation (Stanley et al. 1976). Another DE gene in the head kidney which is highly expressed in lake fish, *leucine-rich repeat containing 17* (*lrrc17*), regulates osteoclasts in mice cells (Kim et al. 2009). The repeated domain of this gene is involved in a variety of protein-protein interactions, including binding to pathogen-associated molecular patterns and surface receptors and thus has been studied in pathogen-host interactions (Kedzierski et al. 2004). Some DE genes with putative immune functions are in contrast more highly expressed in river fish. For example, an uncharacterized protein-coding gene (ENSGACG00000000187) is differentially expressed in both head kidney and spleen, and its sequence is homologous to *NOD-like receptor family CARD domain containing 3* (*NLRC3*). *NLRC3* is a negative regulator of innate immune signaling (Zhang et al. 2014), which inhibits the activity of T cells (Conti et al. 2005) and Toll-like receptor (Schneider et al. 2012). Another DE gene that is highly expressed in river populations in the head kidney is *cub and sushi multiple domains 3* (*csmd3*), reported to be associated with periodontal pathogen colonization in human (Divaris et al. 2012). The putative immune-related function of these candidate habitat-specific genes is consistent with the hypothesis that parasites act as important selective agents driving differentiation between river and lake sticklebacks (Eizaguirre et al. 2012b; Eizaguirre et al. 2011; Feulner et al. 2015; Scharsack et al. 2007; Wegner et al. 2003).

To investigate how differences in parasite load between lake and river populations may be reflected in gene expression in the wild, we compared the set of DE genes with the DE gene sets identified in two previous parasite infection experiments performed on G1 stickleback populations. Despite using

different conditions, sequencing technologies and bioinformatic analyses to identify DE genes, this exercise provides information on immune-related functions of DE genes given their putative role in parasite defense based on experimental studies. The two lab-controlled parasite exposure experiments that we compared our results with used three-spined sticklebacks subjected to infection with parasites that are found in their natural environment: the three parasites *Diplostomum pseudospathaceum*, *Anguillicola crassus*, and *Camallanus lacustris* in a study by Lenz et al. (2013), and D. pseudospathaceum in a separate study by Haase et al. (2014). An independent parasite survey performed on our own transcriptome-sequenced fish (Supplementary Table 1) showed that lake fish have a significantly higher abundance of *Diplostomum sp*. than river fish (negative binomial GLM, z=-4.87, p<0.001, see Supplementary Figure 2), whereas *A. crassus* did not show a habitat-specific pattern (binomial GLM, z=-0.075, p=0.94) and the lake-specific parasite *C. lacustris* (Eizaguirre et al. 2011) was only found in one G1_L fish in our samples. Lenz et al. (2013) assessed gene expression in the head kidney following parasite infection carried out with one of the European population pairs (G1_L and G1_R) used in our study. Among the DE genes found in that study, *methyltransferase like 13* (*mettl13*) was expressed at lower levels in the parasite-challenged fish compared to controls, and in lake versus river individuals after a 2nd parasite infection. In our study, this same gene was also differentially expressed with lower expression in the lake populations in the spleen. These results suggest that *mettl13* expression is down regulated when the fish are challenged with more parasites, for example in lakes versus rivers. *mettl13* is therefore an interesting candidate for mediating a differential expression between lake and river sticklebacks shaped by the contrasting parasite environment. These comparisons to experimental studies demonstrate another way of inferring functional insights of candidate genes, which goes beyond functional annotations based on sequence similarity with model organisms. These transcriptomic results are in line with the hypothesis that parasite-mediated selection contributes to lake and river population differentiation, however it is does not act alone but in interaction with other factors under natural conditions.

Limits of the study

Even though we have been able to gain insight into the role of gene expression in population differentiation, various factors confound the analysis of wild-caught animals. For instance, temporal variation in expression, genetic background differences and stochastic environmental fluctuations introduce variation at the transcriptomic level (Harrison et al. 2012; Lenz 2015). Because our samples are derived from different regions and have been caught at different times of the year, geographical and seasonal factors influenced the observed expression patterns. An important biotic aspect with respect to this study is that fish accumulate parasites from spring to autumn, and their immune system responds differently to early and to late parasite infections (Rohlenová et al. 2011). Furthermore, our study focused on macroparasites, but we acknowledge that there are more pathogens and factors in the natural environment that affect fitness, physiology and immune response. For example, it was found that gut microbiota composition in lake sticklebacks might contribute to shape the genetic polymorphism of MHC class IIb genes (Bolnick et al. 2014), a known genetic basis that vary between fish populations (e.g. Eizaguirre et al. 2011). Hence, microparasites most likely also impact the gene expression of the fish in their natural environments.

In addition, factors like temperature and light condition can vary substantially across geographical regions and seasons. Environmental factors cannot be controlled for sampling on large geographical scale and add noise to the data, reducing the ability to detect habitat-specific patterns. However for each location, parapatric lake and river fish were processed at the same time and alternately dissected, minimizing the variation between lake and river fish within sampling locations. Despite analyzing wild-caught individuals, the majority of our samples showed reasonable correlations between replicated individuals (same habitat, population and sex), resulting in an average Pearson correlation of 0.86. Moreover, including multiple lake-river contrasts can help to overcome some of the variance among wild-caught samples, as it is unlikely that environmental fluctuations would produce habitat-specific expression patterns across multiple individuals and populations by chance. Therefore our results are

conservative estimates of habitat-specific gene expression across the replicated systems.

Having a single population pair from Canada might also affect some results. Since the Canadian populations were rather distinct from the other populations, we also conducted DE analyses only on the three European population pairs for a comparison. However, differential expression between lake and river in the two data sets (with and without the Canadian population pair) were significantly positively correlated and about half of the DE genes are found in both data sets (Supplementary Table 4). Therefore, including one geographically distant population pair from Canada allows identifying habitat-specific patterns on a more global scale. It provides an opportunity to examine which genes show consistent habitat-specific expression patterns in fish across continents, forming a subset of the DE genes from all four population pairs (asterisks in Figure 2).

As we studied the transcriptomic profiles of wild-caught fish, a large number of replication in terms of individuals and populations is required to accommodate environmental variations. This results into trading off sample size and sequencing depth. The Encyclopedia of DNA Elements (ENCODE) consortium recommends 30 million pair-end reads of length > 30 nucleotides, in which 20-25 million reads are mappable to the genome for evaluating transcriptional profiles. In our study, the sequencing depths are generally 5x lower than the recommendation, limiting our ability to detect genes with low expression. When we used a more stringent cutoff to filter out weakly expressed genes (at least 2 reads per million in half of the samples), 10715 genes (compared to 12183 with the original cutoff) in the head kidney and 11012 genes (compared to 12503) in the spleen passed the filtering step. 36 out of 73 DE genes in the head kidney and 58 out of 74 DE genes in the spleen remained with the higher cutoff, suggesting at least half of the detected DE results are robust against the low sequencing depth.

Conclusions and prospects

Despite some intrinsic shortcomings, studying gene expression in wild-caught animals provides a view on differential expression responses caused by both genetic and environmental factors. Our study provides additional evidence that environmental differences, which contrast lakes and rivers and amongst those the distinct parasite community, shape differential gene expression patterns in sticklebacks. We utilize results of previous laboratory-controlled experiments to explain the patterns we detected in the wild. This comparison suggests that amongst other factors the distinct parasite community is most likely an important explanatory factor causing expression differences between habitats. Our results add to previous laboratory results by examining the expression patterns of candidate genes under natural conditions. Those genes identified both here and in previous laboratory studies deserve special attention in potential follow up studies.

Acknowledgements

We thank Derk Wachsmuth for computational assistance and MPI technicians, especially Anja Hasselmeyer for help with lab work. We thank Anika Witten for help with RNA-Seq library preparation and sequencing. We thank Ludovic Mallet for discussions on the study. We thank the International Max Planck Research School for Evolutionary Biology for research support. We also acknowledge the work and generosity of sample collectors including Andrew Hendry, Renaud Kaeuffer, Shahin Muttalib, Caroline Leblond, Noémie Erin, Per Jakobsen, Tom Klepaker and Hendrik Schultz. We thank Sean Rogers and five reviewers for their insightful and constructive comments on a previous version of this manuscript.

Author Contributions

MM, TR and EBB initiated and designed the project. FJJC, CE and MK organized and contributed to the sample collections and dissections. FJJC, IES, MS and PGDF prepared the RNA samples. YH performed quality assessment of the sequencing data, transcriptome mapping, and processed data for analysis. YH, FJJC, MP and PGDF designed the differential expression analyses, and all authors contributed

to discussions on research design and results interpretations. YH drafted the manuscript together with FJJC and PGDF. All authors revised the manuscript.

Data accessibility

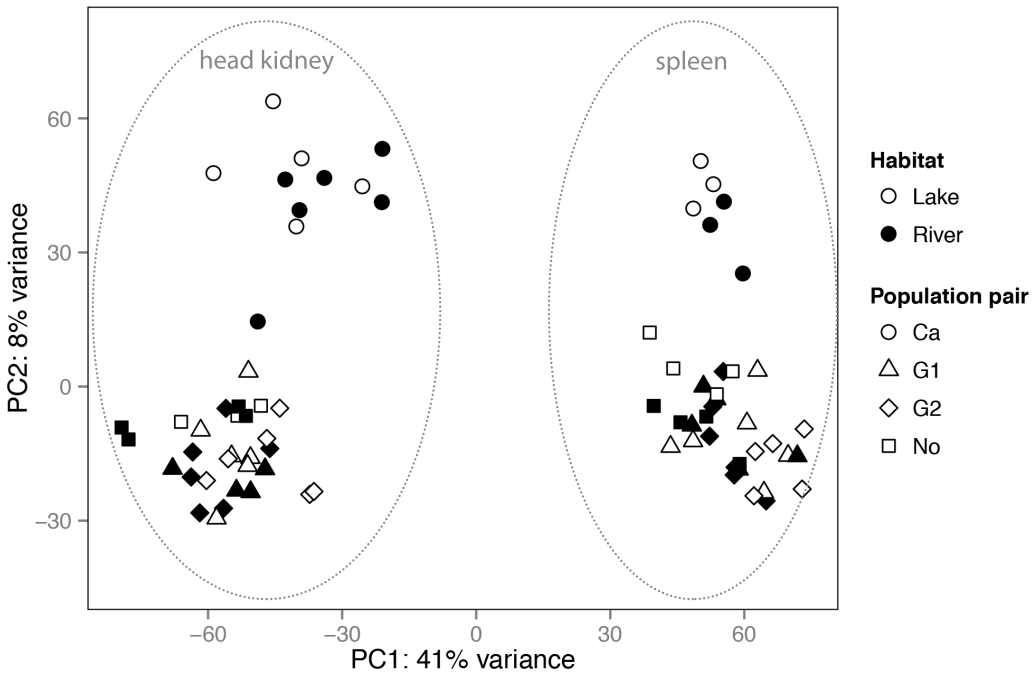
The raw reads of RNA-Seq data (fastq format) and mapping files (bam format) are available through the European Nucleotide Archive (study accession number: PRJEB8677, URL: http://www.ebi.ac.uk/ena/data/view/PRJEB8677). HTSeq read counts, EdgeR results and topGO results are archived in Dryad (doi:10.5061/dryad.hq50s). Morphological and parasite data are included in Table S1.

Table and Figures

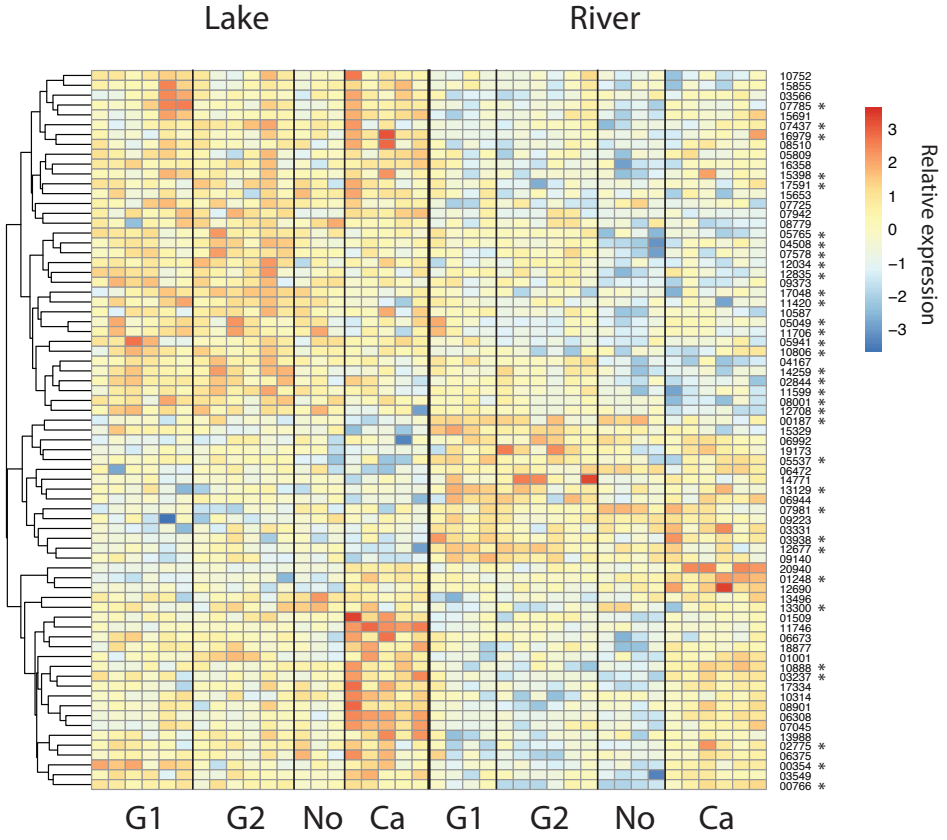
Figures

Fig. 1 Principal component analysis (PCA) of gene expression profiles based on all genes after filtering out weakly expressed genes (See Methods). Head kidney samples and spleen samples are separated along the x-axis, and the Canadian samples are separated along the y-axis. PCA axes explain 41% (x-axis) and 8% (y-axis) of the total variation.

[1] Supplementary materials can be found online: http://onlinelibrary.wiley.com/doi/10.1111/mec.13520/full


head kidney
spleen
PC2: 8% variance
PC1: 41% variance
60
30
0
−30
−60
−30
0
30
60
Habitat
Lake
River
Population pair
Ca
G1
G2
No

(a)


Lake
River
Relative expression
3
2
1
0
−1
−2
−3
G1
G2
No
Ca
G1
G2
No
Ca

(b)

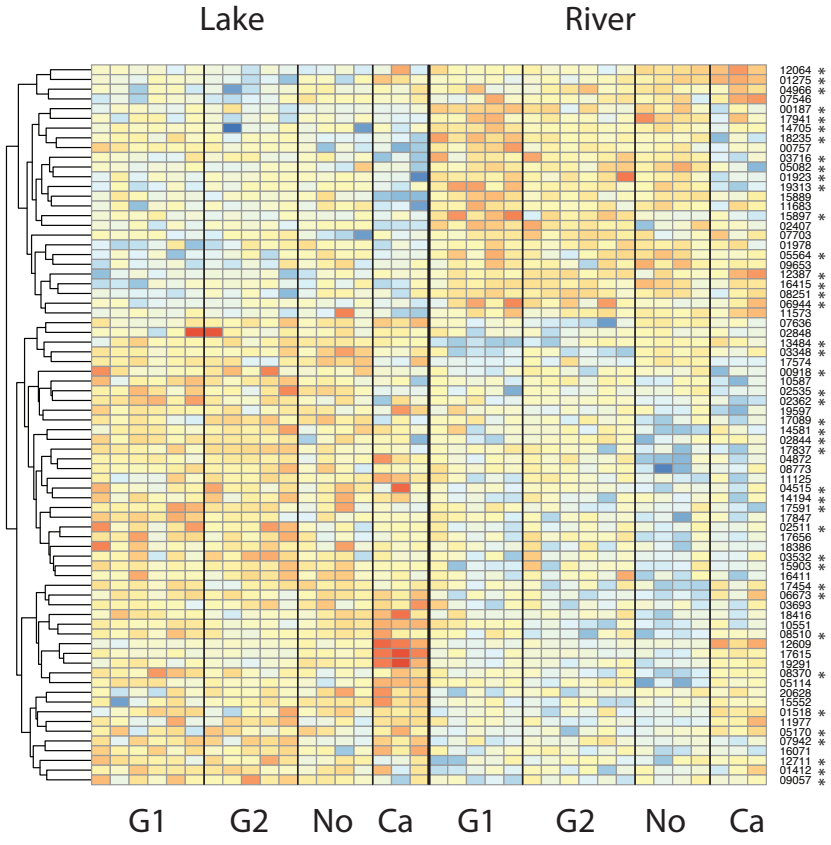

Lake
River
G1
G2
No
Ca
G1
G2
No
Ca

Fig. 2 Heatmaps of DE gene expression profiles among all populations in (a) head kidney and (b) spleen. Each column represents one fish and each row represents one gene. Samples are organized by population affiliation as indicated at the bottom. Genes are clustered based on the similarities of the expression profiles between samples. The color code corresponds to the relative expression intensity, which are the normalized read counts also scaled for each gene's expression intensity (median read count as 0), where red indicates higher expression and blue indicates lower expression. On the right side, the last five digits of the corresponding Ensembl ID (ENSGACG000000XXXXX) are shown. Asterisks indicate genes that were also identified in an analysis of the European populations only (Supplementary Table 4).

Table 1. Summary of sample site information and number of individuals included in the transcriptomic analysis.

Population pair	Location	Habitat	Name	Head kidney	Spleen
G1	Germany	Lake	Großer Ploener See (G1_L)	6	6
		River	Malenter Au (G1_R)	5	5
G2	Germany	Lake	Westensee (G2_L)	6	5
		River	Eider (G2_R)	6	6
No	Norway	Lake	Skogseidvatnet (No_L)	3	4
		River	Orraelva (No_R)	4	4
Ca	Canada	Lake	Misty Lake (Ca_L)	5	3
		River	Misty Stream Inlet (Ca_R)	6	3

Table 2. Differentially expressed genes between all lake and river populations with putative immune functions

Gene ID	Gene name	GO term (biological process)	Tissue	Log fold-change *	FDR
ENSGACG00000001509	*marco macrophage receptor with collagenous structure*	scavenger receptor activity (molecular function)	head kidney	0.73	0.0053
ENSGACG00000016979	*CMKLR1 (2 of 2) chemokine-like receptor 1*	G-protein coupled receptor signaling pathway	head kidney	0.77	0.0070
ENSGACG00000015855	*RAB27A, member RAS oncogene family*	nucleocytoplasmic transport small GTPase mediated signal transduction signal transduction intracellular protein transport	head kidney	0.56	0.026
ENSGACG00000010551	*mst1ra macrophage stimulating 1 receptor a*	protein phosphorylation	spleen	0.89	0.0030
ENSGACG00000012609	*LGALS1 (2 of 3) lectin, galactoside-binding, soluble, 1*	carbohydrate binding (molecular function)	spleen	0.73	0.0038
ENSGACG00000004966	*IRF4 (2 of 2) interferon regulatory factor 4b*	regulation of transcription, DNA-templated	spleen	-0.59	0.028

ENSGACG00000019291	*irak3 interleukin-1 receptor-associated kinase 3*	signal transduction protein phosphorylation	spleen	0.42	0.048
ENSGACG00000001978		antigen processing and presentation immune response	spleen	-1.44	0.048

*: Positive values represent higher expression in lake fish than in river fish and vice versa.

Table 3. Differentially expressed genes between lake and river populations also found as differentially expressed in previous parasite infection studies

Gene ID	Gene name	Comparisons in Lenz *et al.* 2013*	Tissue in Lenz *et al.* 2013	Log fold-change in Lenz *et al.* 2013**	Tissue in this study	Log fold-change in this study ***	FDR in this study
ENSGACG00000011746	*fyco1a FYVE and coiled-coil domain containing 1a*	control vs. infected	head kidney	-3.21	head kidney	0.71	0.0096
ENSGACG00000010806	*sox7 SRY-box containing gene 7*	control vs. infected	head kidney	-2.58	head kidney	0.74	0.038
ENSGACG00000013129	*MRPL49 (2 of 2) mitochondrial ribosomal protein L49*	control vs. infected	head kidney	-2.69	head kidney	-0.91	0.017
ENSGACG00000015653	*lmo1 LIM domain only 1*	control vs. infected	head kidney	-1.20	head kidney	1.13	0.033
ENSGACG00000014705	*mettl13 methyltransferase like 13*	Lake vs. River in 2nd infection; control vs. infected	head kidney	-4.4 and -2.69	spleen	-0.63	0.028
ENSGACG00000011977	*ppdpfa pancreatic progenitor cell differentiation and proliferation factor a*	control vs. infected	head kidney	1.03	spleen	3.17	0.011
ENSGACG00000001923	*n6amt2 N-6 adenine-specific DNA methyltransferase 2*	control vs. infected	head kidney	1.56	spleen	-0.88	0.025
ENSGACG00000004515	*Cfd complement factor D (adipsin)*	control vs. infected	head kidney	-1.48	spleen	1.16	0.00065
ENSGACG00000012609	*LGALS1 (2 of 3) lectin, galactoside-binding, soluble, 1*	control vs. infected	head kidney	-5.09	spleen	0.73	0.0038
ENSGACG00000011683	*slc5a6b solute carrier family 5, member 6*	control vs. infected	head kidney	-2.46	spleen	-0.45	0.045
Gene ID	Gene name	Comparisons in Haase *et al.* 2014*	Tissue in Haase *et al.* 2014.	Log fold-change in Haase *et al.* 2014**	Tissue in this study	Log fold-change in this study ***	FDR
ENSGACG00000003716	*CASQ2 (1 of 2) calsequestrin 2*	control vs. clone XII	head kidney	2.8	spleen	-0.97	0.0080
ENSGACG00000017615	*smox spermine oxidase*	control vs. clone I, control vs. XII and control vs.	head kidney	4.79, 5.81 and 5.11	spleen	0.65	0.011

		clone mix					
ENSGACG00000011977	*ppdpfa pancreatic progenitor cell differentiation and proliferation factor a*	control vs. clone I	head kidney	-4.08	spleen	3.17	0.011
ENSGACG00000004966	*IRF4 (2 of 2) interferon regulatory factor 4b*	control vs. clone I, control vs. XII and control vs. clone mix	head kidney	1.41, 2.26 and 1.79	spleen	-0.59	0.028
ENSGACG00000020628	*angptl5 angiopoietin-like 5*	control vs. clone mix	head kidney	2.83	spleen	0.71	0.028
ENSGACG00000017656	*SVIL (2 of 2) supervillin*	control vs. clone mix	head kidney	2.96	spleen	0.59	0.040
ENSGACG00000008510	*apnl actinoporin-like protein*	control vs. clone mix	gill	-1.20	head kidney and spleen	0.88 and 1.05	0.0093 and 0.0030

*: Comparisons where the genes were previously identified as differentially expressed are indicated. In Lenz *et al.* 2013, DE gene sets between control naïve fish from lake and from river, between twice exposed fish from lake and from river (2nd infection), and between infected fish and control fish were compared to DE gene sets in this study. In Haase *et al.* 2014, DE gene sets between control fish and infected fish with different parasite clones were compared.
For the DE genes Haase *et al.* 2014 identified in gill, only the overlapped DE genes we identified in both head kidney and spleen are shown.
**: In lake-river comparisons, positive log fold-change values represent higher expression in lake fish and vice versa. In control-infection comparisons, positive values represent up-regulation with infection compared to control.
***: Positive log fold-change values represent higher expression in lake fish and vice versa.

Chapter III Copy number divergence contributes to differential gene expression between stickleback ecotypes revealed by combined genome- and transcriptome-wide analyses

Yun Huang*[1], Philine GD Feulner[2,3], Christophe Eizaguirre[4], Tobias L Lenz[1], Erich Bornberg-Bauer[5], Manfred Milinski[1], Thorsten BH Reusch[6], Frédéric JJ Chain[7]

[1] Department of Evolutionary Ecology, Max Planck Institute for Evolutionary Biology, 24306 Plön, Germany

[2]Department of Fish Ecology and Evolution, Centre of Ecology, Evolution and Biogeochemistry, EAWAG Swiss Federal Institute of Aquatic Science and Technology, Seestrasse 79, 6047 Kastanienbaum, Switzerland

[3]Division of Aquatic Ecology and Evolution, Institute of Ecology and Evolution, University of Bern, Baltzerstrasse 6, 3012 Bern, Switzerland

[4] School of Biological and Chemical Sciences, Queen Mary University of London, E1 4NS London, UK.

[5] Institute for Evolution and Biodiversity, Evolutionary Bioinformatics, Westfälische Wilhelms University, 48149 Münster, Germany

[6] Evolutionary Ecology of Marine Fishes, GEOMAR Helmholtz Centre for Ocean Research Kiel, 24105 Kiel, Germany

[7] Department of Biological Sciences, University of Massachusetts Lowell, Lowell MA, 01854, USA

*Correspondence to: yun.huang@evolbio.mpg.de

This chapter is in preparation for submission to *Genome Biology and Evolution*.

Keywords

eCNV, *cis*-regulatory sequence divergence, habitat-specific adaptation, three-spined stickleback

Abstract

Habitat-specific gene expression amongst independent populations suggests a response to the environments potentially shaped by habitat-associated selection, but it is unclear whether or which genetic variants are underlying. Sequence variations in regulatory regions as well as copy number variations (CNVs) both can affect gene expression and potentially contribute to habitat-specific adaptation. Combining genome-wide variant calls of CNVs and SNPs and transcriptome profiles, we investigate the genetic variants associated with habitat-specific gene expression between lake and river populations of the three-spined stickleback (Gasterosteus aculeatus). We did not find noticeable sequence differentiation in putative cis-regulatory regions for genes with habitat-specific expression, but two genes showed habitat-specific patterns both in gene copy numbers and gene expression. These two genes are amongst a total of 135 CNV genes having a positive correlation between gene copy number and gene expression, revealed by individual-based correlation analyses. The correlation between gene copy number and gene expression suggests that for these two genes CNV differentiation between lake and river sticklebacks accompany the habitat-specific gene expression. Taken together, our study highlights copy number variation as a source of genetic variation that can facilitate adaptation to novel environments.

Introduction

Uncovering the genetic basis underlying adaptive phenotypes, which confer increased fitness on organisms, is an ongoing and intense research focus in evolutionary biology (Barrett and Hoekstra 2011). Phenotypes can be modified, for example, by changing amino acid sequences that affect protein structure or by changing regulatory regions that alter the level or spatiotemporal pattern of gene expression. There is a growing body of literature linking adaptive phenotypes in various organisms to gene expression changes, including elongated beaks in cactus finches (Abzhanov, et al. 2006), camouflage pigmentation in deer mice (Linnen, et al. 2009; Mallarino, et al. 2017), and repeated pelvic loss in three-spined sticklebacks (Chan, et al. 2010). Genome-wide analyses of diverging ecotypes or adaptive radiations in several species have also highlighted the importance of regulatory changes in adaptation to different ecological niches (Whitehead and Crawford 2006; Jones, et al. 2012; Brawand, et al. 2014).

Similar to morphological traits, gene expression variation can be used to infer the role of selection in population divergence (Harrison, et al. 2012). Variation in a trait correlated with environmental factors, such as habitat types, suggests it under selection (Endler 1986). Evidence for habitat-shaped gene expression is strong when similar expression patterns have independently arisen in similar habitats (Savolainen, et al. 2013). Albeit such parallel gene expression has been observed in a few cases of diverging ecotypes or adaptive radiated species (Derome, et al. 2006; Pavey, et al. 2011; Manousaki, et al. 2013), the underlying genetic basis of this expression is largely unknown.

Gene expression divergence has been found to have a significant heritable component (Stamatoyannopoulos 2004; Pavey, et al. 2010). One commonly used approach to study the genetic basis underlying gene expression is the mapping of expression quantitative trait loci (eQTL). This method treats gene expression levels as quantitative traits and aims to associate the variation in gene expression to that of genetic variation (Gilad, et al. 2008). eQTL studies identify both *cis*-localized eQTL and *trans*-localized eQTL. Mutations altering the

sequence of *cis*-regulatory elements (CREs) can, for example, affect the binding affinity of transcription factors and subsequent gene expression levels (Wittkopp and Kalay 2011). As CREs have mostly local effects on gene expression compared to *trans*- and thus lower pleiotropy, CREs are presumably favored during population divergence and are primary drivers of gene expression differences between species (Wittkopp, et al. 2008).

An additional type of genetic difference that can alter expression is copy number variation (CNV). CNVs are deletions or duplications of genetic regions that can encompass genes (CNV genes) and impart dosage effects, for example when higher gene copy number leads to increases in gene expression (Haraksingh and Snyder 2013; Gamazon and Stranger 2015). However, copy number changes are not always positively correlated with expression level changes due to various mechanisms, such as compensatory effects via negative feedback in regulation networks or differences in chromatin profiles among copies (Henrichsen, et al. 2009). Nevertheless, CNVs can be associated with putatively adaptive expression changes, such as the duplication of amylase genes in high-starch diet populations of both humans and dogs (Perry, et al. 2007; Axelsson, et al. 2013).

Here we study the genetics of differential gene expression associated with habitat-specific adaptation in the three-spined stickleback (*Gasterosteus aculeatus*). After the last glaciation, marine sticklebacks rapidly and repeatedly colonized different freshwater habitats, resulting in an adaptive radiation composed of a complex of populations and ecotypes (McKinnon and Rundle 2002). Amongst the various habitats colonized, recurrent adaptation to lakes and rivers has given rise to distinct ecotypes (Reusch, et al. 2001; Feulner, et al. 2015). Ecological differences between lakes and rivers include distinct parasite communities, in which lake fish generally suffer from higher parasite pressures than river fish, likely shaping recurrent ecotype differences at both the phenotypic and genetic level (Kalbe, et al. 2002; Eizaguirre, et al. 2011; Feulner, et al. 2015). There is however substantial variation in the patterns of genome-wide differentiation among repeatedly diverged lake and river ecotypes, indicative of local adaptation (Feulner, et al. 2015; Stuart, et al. 2017). Similarly,

differentiation in copy numbers of genes with environmentally-associated functions suggest a role of CNVs in adaptive divergence in sticklebacks (Chain, et al. 2014). At the expression level, experimental infections revealed habitat-specific transcriptional responses to parasite exposure in sticklebacks (Lenz, et al. 2013), and a broad survey of transcriptome profiles among lake and river ecotypes identified 12 immune-related genes amongst 139 genes displaying habitat-specific gene expression (Huang, et al. 2016). Whether there is a relationship between genetic divergence and expression divergence remains to be seen.

In this study, our goal was to identify the genetic variation underlying recurrent expression divergence that putatively contributes to habitat-specific adaptation between lake and river ecotypes in three-spined sticklebacks. We first investigated habitat-specific patterns in genetic variants, both in terms of CRE sequence and of genic CNVs. We then evaluated the correspondence of these habitat-specific patterns with gene expression. We finally evaluated the dosage effects for each CNV gene using the genomes and transcriptomes from the same individuals. By assessing the relationship between CNVs and gene expression, we identified genes whose copy number likely influences differential expression between ecotypes, putatively contributing to habitat-specific adaptation.

Methods

Samples

To study genetic divergence between lake and river stickleback ecotypes and its contribution to expression divergence, we combined a whole genome and a whole transcriptome dataset from a total of eight geographically widespread populations of three-spined sticklebacks that had previously been analyzed separately. The whole-genome sequence dataset (Feulner, et al. 2015, Chain, et al. 2014, EBI Accession no: PRJEB5198) consisted of 48 individuals from 4 parapatric population pairs; 2 independent drainages from Germany (G1 and G2), one from Norway (No), and one from Canada (Ca), with 6 individuals from

each lake (_L) and river (_R) (for detailed information, see Feulner, et al. 2015). The whole-transcriptome dataset (Huang, et al. 2016, EBI Accession no: PRJEB8677) comprised transcriptome sequence data from a subset of the same individuals as above (43 fish, due to suboptimal transcriptome libraries). These transcriptomes had been used to investigate habitat-specific gene expression between lake and river ecotypes among two main immune tissues (head kidney transcriptomes of 40 fish and spleen transcriptomes of 36 fish) (Huang, et al. 2016). The final set of expression profiles consisted of 12,105 genes expressed in the head kidney and 12,451 genes and in the spleen.

Sequence divergence

For each of 19,655 autosomal protein-coding genes, we calculated the sequence divergence between lake and river ecotypes in the 5kb upstream regions. We reasoned that the 5kb upstream regions serve as a proxy for the location of potential *cis*-regulatory elements or "CREs" (Shen, et al. 2012). Single nucleotide polymorphisms (SNPs) identified in these regions were extracted from a previously analysis (Feulner, et al. 2015) and were filtered for a minor allele frequency greater than 0.05 in the four population pairs combined. Sequence divergence between lake and river fish was evaluated using the AMOVA approach implemented in Arlequin (Excoffier and Lischer 2010). The hierarchical structure included 4 lake and 4 river populations represented by 6 individuals each. We calculated the percentage of variance between groups (lake versus river fish) relative to the total variance using the F-statistic "F_{CT}" in Arlequin. For each gene, the F_{CT} for the 5kb upstream region was calculated. We used permutation tests implemented in Arlequin to determine the significance of the F_{CT} values ($p<0.05$). We then required the divergent CRE regions to contain at least one SNP with significant F_{CT} based on locus-by-locus AMOVA (see methods), to ensure that the region is not just on average more divergent but also harbors at least one potentially causal variant. We also calculated F_{CT} of each individual SNP using a locus-by-locus AMOVA analysis with the same hierarchical settings for the upstream regions with significant F_{CT}. This AMOVA approach to detect habitat-specific patterns is more sensitive than methods that scan for significantly divergent regions between each population pair separately

before identifying common divergent regions across pairs (e.g. Feulner, et al. 2015); the habitat-specific patterns detected by the AMOVA approach might not be significant in a given population pair, but significant when comparing all populations together.

Copy number divergence

Similar to the AMOVA approach above, we evaluated copy number divergence between ecotypes across all population pairs together. Utilizing CNV regions identified by Chain et al. (2014), where consensus calls of CNVs were applied by a read depth approach (CNVnator) (Abyzov, et al. 2011) and at least one other approach (paired-end and split-reads; for details see Chain *et al.* 2014), we first identified genes with at least 95% length overlap with the CNV regions. Gene copy number was estimated using CNVnator and rounded to an integer. Genes showing no variation in estimated copy numbers amongst individuals were excluded from copy number divergence analyses. Genes with copy number estimates of zero but with detectable read depth above zero were removed due to possible false deletion calls by CNVnator. A total of 832 autosomal protein-coding genes remained, referred herein as "CNV genes". For each CNV gene, we calculated V_{CT} (Redon, et al. 2006), which evaluates the relative variance in copy number between groups (here lake versus river ecotypes) compared to overall variance within groups, analogous to F_{CT}. V_{CT} was calculated with an ANOVA-based approach, where we treated lake ecotypes and river ecotypes combined from all population pairs as two comparison groups and also accounted for differences between population pair (copy_number ~ ecotypes * population_pair). In this way, V_{CT} is different from previous calculations of pair-wise V_{ST} between each lake and river population pair in Chain *et al.* (2014). To determine how likely each V_{CT} value could be obtained by chance, we randomly permutated the ecotype labels 1000 times. The one-sided p-values were calculated as the fraction of permutated values that exceeded the observed value and were corrected by the Benjamini-Hochberg method for multiple testing (Benjamini & Hochberg 1995). V_{CT} values with corrected p-values smaller than 0.05 were considered as significantly divergent between lake and river ecotypes.

Expression divergence

All of the previously published transcriptome libraries from the sampled populations were included in this study (Huang, et al. 2016; Dryad doi:10.5061/dryad.hq50s). Transcriptome libraries were first normalized using the trimmed mean of M-value (TMM) method (Robinson and Oshlack 2010) across all individuals in EdgeR (Robinson, et al. 2010). Expression levels were estimated as the log of normalized read count per million, following the methods in Huang et al. (2016). Gene expression divergence between ecotypes was then assessed for a total of 12,105 genes from the head kidney and 12,451 from the spleen after removing genes with weak expression (genes with read count per million less than one in at least half of the samples). Huang et al. (2016) employed differential expression (DE) analyses to identify habitat-specific gene expression, which classified genes into two categories: 'significantly differential expressed' versus 'not differentially expressed'.

To complement the binary categorizing of expression divergence by DE analyses, we sought to quantify the extent of expression divergence in a similar way that we estimated F_{CT} and V_{CT}, using the continuous variable P_{CT}, which evaluates the relative variance in expression among populations versus within populations (Antoniazza, et al. 2010). We calculated P_{CT} between lake and river ecotypes from all four population pairs together using an ANOVA-based approach as was done with V_{CT} above, where population pair and sex were accounted for to remove population-specific effects on gene expression and sex-biased effects (following the method from Uebbing, et al. 2016). P_{CT} was calculated for head kidney and spleen separately. Because the calculation of P_{CT} is conceptually equivalent to the calculation of sequence divergence (F_{CT}) and copy number divergence (V_{CT}), the evaluation of expression divergence is made directly comparable to that of genetic divergence (Leinonen, et al. 2013). To make sure P_{CT} complements the previous DE analysis conducted on the same dataset from Huang et al. (2016), we compared the results from the two approaches. We first applied permutation tests to identify genes with significant P_{CT}. The permutation tests and multiple testing corrections for each P_{CT} value were performed in the

same way as V_{CT} (see above). Significant P_{CT} values were determined using a corrected p-value smaller than 0.05. The corrected p-values from P_{CT} analyses were also used for Spearman correlation tests to compare with those from DE analyses, to confirm the consistency between the two methods.

Test for eCNV genes

For each CNV gene, we evaluated the association between gene copy number and expression level across all individuals using a linear mixed effect model. Gene copy number was set as the fixed effect, and the population and sex were random effects (expression ~copy_number + (1|population) + (1|sex)). Tissues were analyzed separately. Benjamini-Hochberg's multiple test correction was applied to the p-values of the fixed effect of copy number. Genes with corrected p-values smaller than 0.05 were considered as "eCNV genes", having statistically significant correlations between copy number and expression.

Statistical analyses

For gene sets with significant F_{CT}, significant V_{CT}, and the joint set of eCNV genes from either tissues, the enrichment in gene ontology (GO) terms were tested with topGO (Alexa and Rahnenfuhrer 2016), based on Fisher's exact tests applying Benjamini-Hochberg's multiple-test correction. The gene sets were compared to the all genes annotated in the stickleback reference genome, if not otherwise stated. Overrepresented GO terms were those with a corrected p-values (FDR) smaller than 0.05. All statistical analyses were carried out using the package R version 3.0.2 (R Development Core Team 2011) unless otherwise indicated.

Results

Few genes with habitat specific sequence divergence in cis-regulatory regions

Our goal was to evaluate the relationship between genetic divergence and gene expression divergence between replicated pairs of lake and river three-spined stickleback ecotypes. We first calculated sequence divergence in putative CREs (5kb upstream regions) of protein coding genes between fish from the two

contrasting habitats. We found a total of 10 out of about 20 thousand autosomal protein-coding genes were significantly diverged between fish from the different habitats based on the F_{CT} of the CRE regions ($p<0.05$, permutation test, Figure 1a, Table 1). Based on the F_{CT} of individual SNPs in these 10 gene CREs, 7 genes had multiple significantly diverged SNPs between ecotypes (($p<0.05$, permutation test), probably forming a divergent haplotype. Four of the 10 divergent CRE regions have functions related to signal transduction (Table 1), but no functional groups were enriched.

Copy number divergence in dozens of genes

Besides sequence divergence in the CRE regions, we evaluated divergence in gene copy numbers between ecotypes, which could also result in gene expression changes. Gene copy number divergence between lake and river fish was performed across all 4 population pairs together, rather than between each population pair separately (Chain, et al. 2014). Out of a total of 19782 protein-coding autosomal genes in stickleback genome, we focused on 832 CNV genes detected among our samples (4.21%), for which we calculated the copy number divergence between ecotypes (V_{CT}, Table S1). A total of 4.3% of the CNV genes (36 genes) had a significant V_{CT} (FDR<0.05, permutation test, Figure 1b, Table 2), with V_{CT} values ranging from 0.117 to 0.578. Twenty-three of these genes have higher average copy numbers in lake ecotypes than in river ecotypes, and the reciprocal was true for 13 genes. Twenty-five of the 36 significant V_{CT} genes were from unique CNV regions while other 11 were encompassed in 5 CNV regions. The CNV genes from same CNV regions showed correlated copy number patterns across individuals, suggesting linkage by same duplication or deletion events. The 36 significant V_{CT} genes were enriched with the gene ontologies (GO) related to protein glycosylation (biological process) compared to the genomic background, which was also enriched in the whole CNV gene set (Chain et al. 2014). The 36 significant V_{CT} genes showed no functional enrichment compared to other CNV genes, indicating no particular functions were preferentially diverged within the CNV gene set.

Quantitative expression divergence as a complement to differential expression analyses

We sought to evaluate the extent of expression divergence (P_{CT}) in a similar quantitative fashion as F_{CT} and V_{CT}, e.g. gene expression variances within versus between fish ecotypes. We identified 115 and 88 genes with significant P_{CT} in the head kidney and spleen, respectively (FDR<0.05, Figure 1c). In comparison, we had previously identified 73 and 74 DE genes in the head kidney and spleen, respectively. Half of the DE genes had a significant P_{CT}. The DE genes as a whole exhibited significantly higher P_{CT} values than non-DE genes ($p<0.001$, Wilcoxon rank sum test). In addition, genes having higher P_{CT} also tend to have higher magnitude of gene expression differences (absolute values of log fold changes) between lake and river fish (rho=0.733 in head kidney and rho=0.755 in spleen, $p<0.001$ in both cases, Spearman correlation test), confirming P_{CT} as a sensitive quantitative estimation of expression divergence, complementing DE analyses.

Little genome-wide correlation between genetic divergence and expression divergence

We measured the correlation between F_{CT} in CRE regions and P_{CT} across all genes to test whether CRE sequence divergence is positively correlated with expression divergence genome-wide. We found that F_{CT} and P_{CT} are barely positively correlated (rho=0.017, p=0.03, n=12,057 in head kidney and rho=0.012, p=0.09, n=12,400 in spleen, Spearman correlation test). Very weak positive or absent correlations were also found when analyzing each population pair separately. Amongst the 10 genes with significant sequence divergence (see above), 5 have no expression information in our transcriptome data, and the other 5 were not DE genes and had P_{CT} values below 0.1 (Table 1).

Similarly, we found little to no positive correlation between V_{CT} and P_{CT} overall (rho=0.064, p=0.064 for head kidney; rho=0.166, $p<0.001$ for spleen; Spearman rank correlation). In other words, the level of divergence in copy numbers does not strongly predict the level of expression divergence. Nevertheless, the 9 CNV genes that had significant P_{CT} had higher mean V_{CT} compared to other CNV genes (p=0.016, Wilcoxon rank sum test), two of which also had significant V_{CT}. One

gene is *cathepsin A* (ENSGACG00000015897*)*, which had the highest V_{CT} (0.58) amongst all CNV genes and was previously identified as the most differentiated CNV gene between ecotypes in two German population pairs (Chain, et al. 2014). Here we show that the copy number divergence of this gene is accompanied by significant gene expression divergence (Figure 2). The signal is dominated by the two German population pairs in which river individuals had both higher copy numbers and higher expression levels than lake individuals.

The second gene exhibiting both a significant V_{CT} and P_{CT} value was *GTPase, IMAP family member 7* (*GIMAP7*, ENSGACG00000018877) with a V_{ST} of 0.35. This gene has high V_{CT} values in 3 population pairs and a moderate value in the fourth population pair (G2), with higher copy numbers and higher expression in lake ecotypes except in G2 population pair (Figure 2). It appears plausible that the copy number divergence (V_{CT}) of these two genes represents a genetic basis underlying their gene expression divergence.

CNV genes associated with expression variation through dosage effects

To evaluate whether CNVs influence gene expression levels regardless of ecotype divergence, we tested the association between gene copy number and gene expression on an individual-by-individual basis. Among 350 expressed CNV genes, 140 had a significant association between gene copy number and gene expression in at least one of the two immune tissues (corrected p values< 0.05). Five of these genes had a significant negative correlation between copy number and expression level, *WBP1*(*WW domain binding protein 1*, ENSGACG00000000318) and *slc47a1* (*solute carrier family 47, member 1*, ENSGACG00000020614) and two uncharacterized genes (ENSGACG00000020469 and ENSGACG00000012806) in head kidney samples and *cyp3c1* (*cytochrome P450 family 3 subfamily A member 43*, ENSGACG00000010952) in spleen samples. The other 135 genes with a positive correlation were considered as “eCNV genes” in at least one of the two tissues. Sixty were eCNV genes in both the head kidney and the spleen (out of 117 CNV genes that were expressed in both tissues). The GO enrichment analyses of the eCNV genes against the whole stickleback gene set showed that they were

enriched for antigen processing and presentation (GO:0019882, with 4 of 32 genes), and MHC protein complex (GO:0042611 with 4 of 31 genes). Although these functional categories are already known to be enriched for CNV genes in sticklebacks (Chain et al. 2014), here we show that they also are enriched for an association with dosage effects, i.e. copy number variation might have functional consequences at the phenotypic level by altering the number of gene transcripts. However, when compared to other CNV genes, eCNV genes did not show any GO term enrichments, suggesting that no particular functions were preferentially expressed with dosage effects within the CNV gene set.

The two genes with both significant V_{CT} and P_{CT} (described above) were also eCNV genes (Figure 2). The gene *cathepsin A* had significant P_{CT} and DE identified in spleen and copy numbers highly correlated with the gene expression in both tissue types (FDR<0.001, Figure 2a). The gene *GIMAP7* had significant P_{CT} and DE identified in head kidney and copy numbers highly correlated with the gene expression also in both tissue types (FDR=0.0074 in head kidney and FDR<0.001 in spleen, Figure 2f). The correlation between gene expression levels and gene copy numbers of these two genes strongly suggests the gene copy number changes as a genetic mechanism underling expression divergence.

Discussion

In this study we combined the analysis of genetic variation and gene expression among individuals and their divergence patterns between ecotypes to better understand the genetics of habitat-specific adaptation. Overall sequence divergence and copy number divergence were only weakly or not correlated with gene expression divergence, generally explaining little variation in genome-wide expression divergence. However, certain genes were found with strong associations between genetic divergence and expression divergence. In particular, we provide evidence that copy number divergence likely shapes gene expression divergence of at least two genes between recently diverged ecotypes through a dosage effect. This supports that CNVs can drive habitat-specific gene expression and potentially facilitate adaptive evolution.

Based on the notion that cis-regulatory changes make a major contribution to phenotypic divergence between populations or species (Wray 2007; Wittkopp and Kalay 2011), we rationalized that higher sequence divergence in CREs would lead to higher divergence in gene expression. Although studies found a major contribution of *cis*-eQTLs to expression variation and expression divergence between ecotypes in sticklebacks (Ishikawa, et al. 2017; Pritchard, et al. 2017), we found little to no genome-wide correlation between putative CRE sequence divergence and expression divergence, which is similar to studies in whitefish and in flycatcher (Renaut, et al. 2012; Uebbing, et al. 2016). Zhao et al. found highly differentiated SNPs enriched in cis-regulatory regions of DE genes between low and high latitude populations of *Drosophila*, but also failed to find a correlation between the magnitude of genetic differentiation and that of expression differentiation (Zhao, et al. 2015). These studies, together with ours, investigate the relationship between CRE and expression divergence in closely related taxa. The ecotypes used in this study were sampled from geographically distant populations and have diverged differently in genomic regions (Feulner, et al. 2015), resulting in the majority of CRE regions with non differentiation (zero or low F_{CT}) with a few exceptions. The non or low differentiation in most CRE regions might leave subtle impacts on expression divergence. On the other hand, plastic responses regulated by the organisms' innate ability probably also contribute to expression divergence (Gibson 2008), masking the effects by the genetic differences. Other explanations include *trans*-regulatory factors that interact with CREs (Metzger, et al. 2017), non-additive effects of genetic loci underlying expression (Merilä and Crnokrak 2001). It is also possible that the influence of CRE sequence divergence on gene expression divergence is manifested in a time and tissue other than what was captured in our sequenced transcriptomes.

Though CNVs interfere with much fewer genes than SNPs in the CREs, making their relative impacts on expression divergence not comparable, up to 16% of variation in expression divergence P_{CT} is explained by V_{CT}, which is qualitatively higher than F_{CT}. The 16% of expression variation explained by P_{CT} reflects a

non-negligible impact of V_{CT} on P_{CT}. This is consistent with previous suggestions that genomic changes modifying the number of copies of a gene have a greater impact on gene expression than sequence differences (Sudmant, et al. 2015; Huddleston and Eichler 2016). Though large proportion of the variation in P_{CT} is not explained by V_{CT}, which again could be due to the same reason as for F_{CT}, we suspect that a subsets of CNV genes should contribute the expression divergence in a fashion of a positive relationship between gene dosage and expression, especially for protein-coding genes (Conrad and Antonarakis 2007). We found that 135 (38.6%) of all CNV genes showing positive correlations between copy number changes and expression changes, referred to as "eCNVs". Similar numbers of genes showing associations between expression changes and CNVs have been reported in different organisms despite using different approaches to identify expression-associated CNV genes (Stranger, et al. 2007; Schlattl, et al. 2011). In *Drosophila*, about half of gene duplications are associated with significant increases in expression. And these expression-associated gene duplications are enriched among low- and high-frequency duplications, suggesting either deleterious or adaptive roles from dosage effects (Cardoso-Moreira, et al. 2016). Although the majority of CNVs with dosage effects is predicted to be deleterious (Veitia 2002, 2005) and eventually purged by selection (Rice and McLysaght 2017), under the right circumstances CNVs can facilitate adaptation via dosage effects on gene expression and further on external physiological or morphological phenotypes (Iskow, et al. 2012).

Previous studies have documented parallel evolution of CNVs between low and high latitude populations in *Drosophila* (Schrider, et al. 2016) and between marine and freshwater populations of sticklebacks (Hirase, et al. 2014), suggesting the contribution of CNVs in adaptation to different environments. In this study we investigated the associations between CNVs and gene expression changes that could have an adaptive role in the divergence between lake and river stickleback ecotypes. The divergent signals both in gene copy number (V_{CT}) and in gene expression (P_{CT}), and the correlation between copy numbers and gene expression (eCNV) serve as three pillars to support that the gene copy number changes contribute to adaptation. That is, compelling evidence for

habitat-specific selection is inferred when a gene has different copy numbers between ecotypes, different gene expression between ecotypes, and has higher expression in individuals with higher copy numbers. We detected two genes displaying all of these three signals, which are strong candidates for adaptive copy number changes. One gene is *cathepsin A,* which had the highest V_{CT} amongst all CNV genes and higher copy number in river populations that could be favored due to dosage effects. This gene codes for a protein that plays an important role in processing endogenous bioactive peptides (Timur, et al. 2016) and muscle metabolism (González-Prendes, et al. 2017). Its isoforms CTS L and S have been shown to have roles in MHC class II antigen presentation (Hsing and Rudensky 2005), which suggests that the copy number divergence of the gene might contribute to differential immune response in the two ecological environments. Another putatively adaptive CNV gene is *GIMAP7*, which has on average higher copy number and correlated higher gene expression in lake populations. *GIMAP7* is a GTPase gene and contains domain AIG1-type G, which has immunity-associated functions that is conserved from plants to vertebrates (Krücken, et al. 2004; Schwefel, et al. 2010). The immune-related functions of these two genes add to previous findings that CNVs act as an important type of genetic variation to maximize the host innate and adaptive response (Chain, et al. 2014; Machado and Ottolini 2015). An increased copy number and the correspondingly increased expression may promote the defense to different parasite communities in lakes and rivers, facilitating adaptation to different environments.

By combining genome and transcriptome data from the same individuals, we brought together signatures at the genetic level and transcription level to evaluate the genome-wide associations between the two, and identified genetic variants underlying gene expression that likely contribute to ecological adaptation. We report evidence of CNVs acting as a genetic mechanism underlying gene expression divergence between ecotypes. Though CNVs are mostly conceived as deleterious mutations, our findings highlight CNVs with their possible contributions to adaptive evolution.

Acknowledgments

We thank the International Max Planck Research School for Evolutionary Biology for research support. We thank Prof. Tal Dagan for discussions on the study. We thank Derk Wachsmuth for computational assistance.

M.M., T.R. and E.B-B. initiated and designed the project. Y.H., F.J.J.C., and P.G.D.F. designed the analyses. Y.H. performed the analyses, and all authors contributed to discussions and interpretation of the results. Y.H. drafted the manuscript together with F.J.J.C. and P.G.D.F. All authors revised the manuscript.

Figures

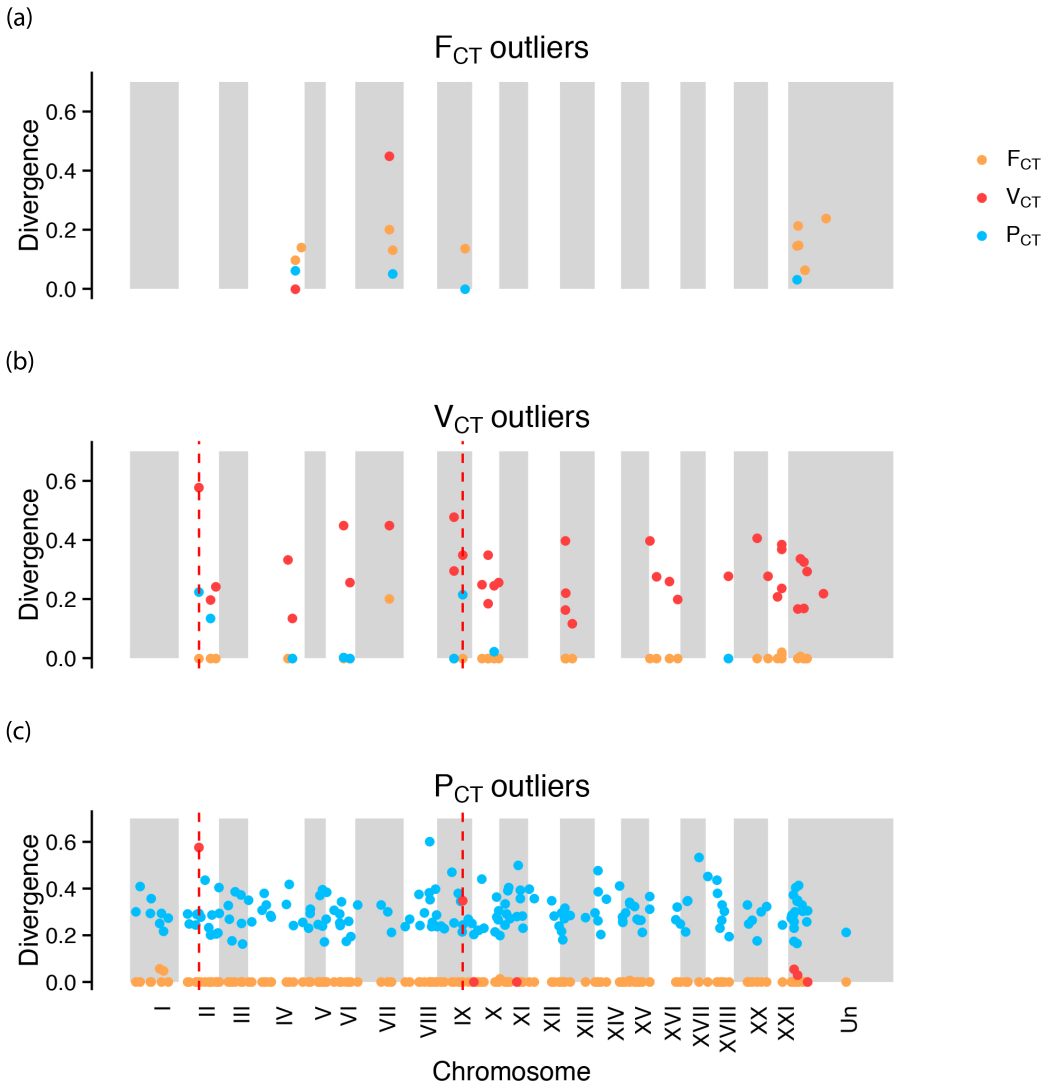


Figure 1. Sequence divergence in 5kb upstream regions (F_{CT}), copy number divergence (V_{CT}) and expression divergence (P_{CT}) of (a) F_{CT} outliers, (b) V_{CT} outliers and (c) P_{CT} outliers along chromosome locations. For each outlier gene, all different divergence estimates of the gene are shown in yellow (F_{CT}), red (V_{CT}) and blue (P_{CT}), when the estimates are applicable[1]. For P_{CT} outliers in both tissue types and for F_{CT} and V_{CT} outliers having P_{CT} values in both tissue types, the average P_{CT} are shown. Twenty linkage groups of the stickleback genome

[1] For each gene, F_{CT} is applicable when a gene compasses SNPs in the 5kb upstream region; V_{CT} is applicable when the gene is a CNV gene; P_{CT} is applicable when the gene is non weakly expressed in either or both tissue types.

representing autosomes together with unplaced scaffolds (Un) are shown and alternated in white and grey backgrounds. The two dashed vertical lines in (b) and (c) indicate the two same genes with both significant V_{CT} and P_{CT}.

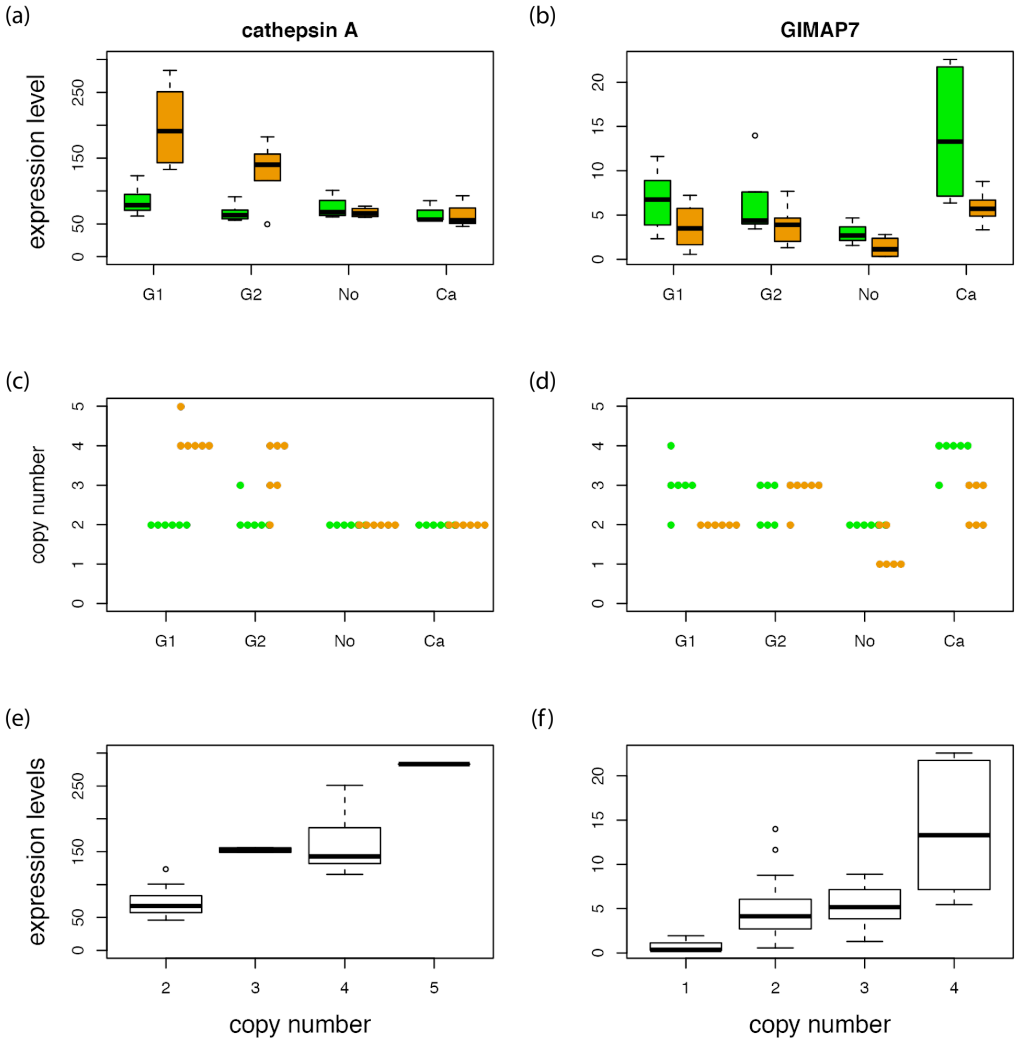


Figure 2. Gene expression level and gene copy number of two candidate DE genes with significant V_{CT} among all populations. Habitat-specific expression patterns of (a) *cathepsin A* and (b) *GIMAP7* across all populations. Y-axes indicate expression levels in normalized read counts. Green boxes are lake populations and yellow boxes are river populations. Habitat specific patterns of gene copy number of (c) *cathepsin A* and (d) *GIMAP7*. Green dots are lake individuals and yellow dots are river individuals. The association between gene expression levels and gene copy numbers of (e) *cathepsin A* and (f) *GIMAP7*.

Tables

Table 1. Genes with significant sequence divergence between lake and river ecotypes in 5kb upstream regions.

Gene ID	Gene name	GO function	F_{CT}	Total no. of SNPs	No. of sig. SNPs	F_{CT} for sig. SNPs	P_{CT} *	
							HK	SP
ENSGACG00000000400	novel gene	unkown	0.240	5	2	0.40, 0.46	NA	NA
ENSGACG00000000842	novel gene	unkown	0.150	37	5	0.36-0.61	NA	NA
ENSGACG00000000998	novel gene	unkown	0.150	8	2	0.34, 0.40	NA	0.031
ENSGACG00000001347	novel gene	zinc ion binding, metal ion binding	0.060	30	0	-	NA	NA
ENSGACG00000015690	novel gene	signal transduction, 3',5'-cyclic-nucleotide phosphodiesterase activity, phosphoric diester hydrolase activity, hydrolase activity, metal ion binding	0.210	3	1	0.320	NA	NA
ENSGACG00000019256	signal recognition particle 68 (srp68)	SRP-dependent cotranslational protein targeting to membrane, 7S RNA binding, endoplasmic reticulum signal peptide binding, signal recognition particle	0.140	31	4	0.38-0.44	0.000	0.000

		binding						
ENSGACG00000019615	novel gene	Proteolysis, serine-type endopeptidase activity	0.100	5	1	0.270	0.062	NA
ENSGACG00000019849	novel gene	unkown	0.140	42	9	0.34-0.49	NA	NA
ENSGACG00000020534	novel gene	signal transduction, G-protein coupled receptor signaling pathway, sensory perception of smell, response to stimulus, olfactory receptor activity,	0.200	43	5	0.32-0.32	NA	NA
ENSGACG00000020630	Rho GTPase activating protein 42a (arhgap 42a)	signal transduction	0.130	40	13	0.23-0.36	0.087	0.016

*: NAs indicate non- or weak expression.

Table 2. Genes with significant divergence in gene copy numbers (V_{CT}) between lake and river ecotypes

Gene ID	Gene name	GO function	V_{CT}					P_{CT} *	
			All **	G1	G2	No	Ca	HK	SP
ENSGACG00000000238	ST3 beta-galactoside alpha-2,3-sialyltransferase 1 (ST3GAL1, 1 of 7)	transferase activity, transferring glycosyl groups,carbohydrate metabolic process,cellular protein modification process,protein glycosylation,biosynthetic process,molecular_function,sialyltransferase activity,biological_process	0.168	0.263	0.105	0.418	0.018	NA	NA
ENSGACG00000000240	ST3 beta-galactoside alpha-2,3-sialyltransferase 1 (ST3GAL1, 2 of 7)	transferase activity, transferring glycosyl groups,carbohydrate metabolic process,cellular protein modification process,protein glycosylation,biosynthetic process,molecular_function,sialyltransferase activity,biological_process	0.326	0.495	0.000	0.308	0.374	NA	NA
ENSGACG00000000408	novel gene	unkown	0.218	0.256	0.546	0.043	0.613	NA	NA
ENSGACG00000000857	novel gene	extracellular region,cellular_component	0.293	0.102	0.000	0.044	0.746	NA	NA
ENSGACG00000001537	novel gene	unkown	0.397	0.267	0.266	0.708	0.386	NA	NA
ENSGACG00000000001	novel gene	unkown	0.278	0.049	0.014	0.676	0.731	NA	NA

645									
ENSGACG00000001748	novel gene	enzyme regulator activity,extracellular region,endopeptidase inhibitor activity,molecular_function,cellular_component	0.336	0.323	0.162	0.060	0.494	NA	NA
ENSGACG00000002209	novel gene	G-protein coupled receptor activity,G-protein coupled receptor signaling pathway,signal transducer activity,signal transduction,molecular_function,cellular_component,biological_process,integral component of membrane	0.275	0.259	0.208	0.667	0.000	NA	NA
ENSGACG00000002473	novel gene	unkown	0.208	0.285	0.136	0.445	0.389	NA	NA
ENSGACG00000003404	novel gene	unkown	0.237	0.735	0.011	0.242	0.000	NA	NA
ENSGACG00000003405	novel gene	unkown	0.368	0.892	0.563	0.167	0.827	NA	NA
ENSGACG00000003407	novel gene	unkown	0.384	0.856	0.461	0.122	0.693	NA	NA
ENSGACG00000003969	novel gene	unkown	0.250	0.314	0.222	0.785	0.316	NA	NA

ENSGACG00000005317	novel gene	unkown	0.163	0.447	0.123	0.003	0.650	NA	NA
ENSGACG00000005319	novel gene	unkown	0.397	0.730	0.000	0.068	0.765	NA	NA
ENSGACG00000005345	novel gene	cellular_component,integral component of membrane	0.220	0.771	0.391	0.000	0.482	NA	NA
ENSGACG00000006218	transient receptor potential cation channel, subfamily M, member 2 (trpm2)	ion channel activity,transport,ion transport,transmembrane transport,transmembrane transporter activity,molecular_function,cellular_component,membrane,biological_process	0.260	0.000	0.000	0.000	0.811	NA	NA
ENSGACG00000006431	novel gene	transferase activity, transferring glycosyl groups,carbohydrate metabolic process,cellular protein modification process,protein glycosylation,biosynthetic process,molecular_function,cellular_component,membrane,fucosyltransferase activity,biological_process	0.349	0.674	0.136	0.850	0.025	NA	NA
ENSGACG00000006432	novel gene	transferase activity, transferring glycosyl groups,carbohydrate metabolic process,cellular protein modification process,protein glycosylation,biosynthetic process,molecular_function,cellular_component,membrane,fucosyltransferas	0.185	0.586	0.204	0.391	0.058	NA	NA

		e activity,biological_process							
ENSGACG00000007399	novel gene	unkown	0.117	0.396	0.000	0.000	0.000	0.000	NA
ENSGACG00000008264	novel gene	unkown	0.245	0.036	0.460	0.682	0.430	0.000	0.045
ENSGACG00000008305	novel gene	G-protein coupled receptor activity,G-protein coupled receptor signaling pathway,signal transducer activity,signal transduction,molecular_function,cellular_component,biological_process,integral component of membrane	0.200	0.123	0.650	0.021	0.000	0.000	NA
ENSGACG00000008985	novel gene	unkown	0.448	0.000	0.067	0.000	0.833	0.004	0.000
ENSGACG00000009880	novel gene	unkown	0.257	0.368	0.123	0.633	0.000	0.000	NA

ENSGACG00000010952	cytochrome P450 family 3 subfamily A member 43 (CYP3A43)	oxidation-reduction process,oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen,heme binding,oxidoreductase activity,ion binding,molecular_function,iron ion binding,biological_process	0.257	0.521	0.010	0.011	0.674	0.000	0.000
ENSGACG00000012073	novel gene	unkown	0.278	0.494	0.371	0.000	0.783	NA	0.000
ENSGACG00000012354	novel gene	unkown	0.406	0.165	0.000	0.217	0.837	NA	NA
ENSGACG00000015897	cathepsin A	peptidase activity,proteolysis,molecular_function,serine-type carboxypeptidase activity,biological_process	0.578	0.962	0.510	0.000	0.000	0.159	0.289
ENSGACG00000016770	novel gene	hydrolase activity,small molecule metabolic process,molecular_function,dUTP metabolic process,cellular nitrogen compound metabolic process,dUTP diphosphatase activity,biological_process	0.197	0.445	0.285	0.000	0.000	0.120	0.148
ENSGACG00000017259	novel gene	unkown	0.242	0.083	0.002	0.183	0.707	NA	NA

ENSGACG00000018037	LON peptidase N-terminal domain and ring finger 1, like (lonrf1l)	ATP-dependent peptidase activity,cellular protein modification process,peptidase activity,proteolysis,protein binding,ion binding,ATPase activity,molecular_function,ubiquitin-protein transferase activity,zinc ion binding,metal ion binding,biological_process,protein ubiquitination	0.478	0.784	0.324	0.000	0.000	0.000	0.000
ENSGACG00000018047	novel gene	ion binding,molecular_function,calcium ion binding	0.296	0.764	0.264	0.000	0.000	NA	NA
ENSGACG00000018877	GTPase, IMAP family member 7 (GIMAP7)	ion binding,GTP binding,molecular_function	0.348	0.533	0.233	0.637	0.701	0.245	0.1844
ENSGACG00000019238	novel gene	extracellular region,cellular_component	0.334	0.014	0.263	0.376	0.817	NA	NA
ENSGACG00000019508	neurexophilin and PC-esterase domain family, member 3 (nxpe3)	unkown	0.134	0.000	0.268	0.194	0.000	0.000	0.000
ENSGACG00000020534	novel gene	G-protein coupled receptor activity,G-protein coupled receptor signaling pathway,signal transducer activity,signal transduction,molecular_function,cellular_component,biological_process,integral component of membrane	0.448	0.573	0.666	0.000	0.695	NA	NA

*: P_{CT} between ecotypes from all populations. NAs in P_{CT} columns indicate non- or weak expression. P_{CT} in red are significant.
**: All means V_{CT} between ecotypes from all populations (ANOVA-based). The other V_{CT} columns contain V_{CT} between parapatric lake and river populations

Conclusion

This thesis addresses genetic and transcriptional parallel evolution across five lake and river population pairs of three-spined sticklebacks spanning two continents. Alongside with many ecological differences between lake and river environments, sticklebacks bear differential parasite pressure between lakes and rivers, a factor that has a strong impact on the fitness. Lake sticklebacks harbor higher parasite loads than river sticklebacks, a consistent pattern across continents. Parasites are conceived as an important selective force driving the lake and river sticklebacks to diverge as two distinct ecotypes. The recurrent lake and river ecotypes differ in their immune competence and allele pools of major histocompatibility complex, likely due to adaptation to the differential parasite pressures between lake and river environments (Scharsack et al. 2007; Eizaguirre et al. 2011). This thesis focuses on genetic changes and gene expression differences to understand the genetic basis of parallel evolution in freshwater sticklebacks.

The genome scan approach detects little parallelism on the genetic level (Chapter I), while transcriptome profiling detects 139 genes with habitat-specific expression patterns (Chapter II). The little genetic parallelism detected in Chapter I suggests that, although under apparently similar selection regimes (habitat-associated selection), evolution at the genetic level follows idiosyncratic trajectories and is infrequently repeatable due to contingent events. This lack of parallelism is in contrast to 139 genes with the habitat-specific patterns on the gene expression level in Chapter II, which represent parallelism on a phenotypic level. A subset of the habitat-specific expression genes are immune genes or involved in responses to parasite infection in previous studies (Lenz et al. 2012; Haas et al. 2014), reinforcing the important role of parasite-mediated selection in shaping habitat-specific adaptation in lake and river fish. It remains possible that the habitat-specific expression patterns are due to plastic response of sticklebacks to the environment differences between lake and river habitats. Alternatively, as the lake and river sticklebacks have adapted to their local environments, we speculate that divergent selection between lake and river

habitats may have acted upon genetic changes, shaping the habitat-specific expression patterns.

To reconcile the discrepancy between low genetic parallelism (Chapter I) and habitat-specific expression patterns found on hundred genes (Chapter II), the third chapter explores the association between habitat-specific gene expression patterns with genetic variants. Here, habitat-specific gene expression patterns are quantified as expression divergence between ecotypes. Expression divergence on the genome-wide scale is not correlated with sequence divergence in cis-regulatory region or gene copy number divergence, suggesting a complex relationship between genetic divergence and expression divergence. Nevertheless, two genes with habitat-specific gene expression are amongst the genes with highest copy number divergence between lake and river ecotypes. The analysis of covariance between gene expression and copy number variation on a gene-by-gene basis revealed these two genes amongst a total of 135 genes showing dosage effects of gene copy number changes on gene expression, strongly suggesting that copy number of these two genes are the underlying genetic control for the habitat-specific expression. These two genes provide strong candidates that the copy number variants being repeatedly selected due to dosage effects on gene expression. Given these findings found across independent populations, we propose that copy number variation as a genetic source promoting adaptation to novel environments.

Taken together, my PhD work combines genome and transcriptome analyses of three-spined sticklebacks from replicated lake and river population pairs to better understand the genetic basis of parallel evolution. We identify a number of genes with habitat-specific expression patterns representing phenotypic parallelism, while a lesser extent of parallelism is found at the genetic level. These results suggest complex genetic mechanisms underlying parallel phenotypes, with plastic responses to environments potentially also playing a role. Nevertheless, when the genetic basis for a phenotype is of a large effect, such as the dosage effect of copy number variations on gene expression patterns, the parallelism on the genetic level is also present. In this case, the parallelism on

both the genetic level and the phenotypic (gene expression) level together reinforces that the parallel genotypes and the corresponding phenotypes have been selectively favored during adaptation to the habitats. Last but not least, the three chapters identify sets of genes with parallel divergence either at the genetic level or at the expression level or both, providing candidate genes for follow-up studies to better understand the genetic basis of adaptation.

Bibliography

Abyzov A, Urban AE, Snyder M, Gerstein M (2011) CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Research 21:974-984.

Abzhanov A, Kuo WP, Hartmann C, Grant BR, Grant PR, Tabin CJ. (2006) The calmodulin pathway and evolution of elongated beak morphology in Darwin's finches. Nature 442:563-567.

Albertson RC, Streelman JT, Kocher TD, Yelick PC (2005) Integration and evolution of the cichlid mandible: the molecular basis of alternate feeding strategies. Proceedings of the National Academy of Sciences of the United States of America 102, 16287-16292.

Alexa A, Rahnenfuhrer J (2010) topGO: Enrichment analysis for Gene Ontology. R package version 2.18.0. .

Alexa A, Rahnenführer J, Lengauer T (2006) Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. Bioinformatics 22, 1600-1607.

Alexa A, Rahnenfuhrer J. (2016) topGO: Enrichment Analysis for Gene Ontology.

Alvarez M, Schrey AW, Richards CL. (2015). Ten years of transcriptomics in wild populations: what have we learned about their ecology and evolution? Molecular Ecology 24:710-725.

Anders S, McCarthy DJ, Chen Y, et al. (2013) Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. Nature Protocols 8, 1765-1786.

Anders S, Pyl PT, Huber W (2014) HTSeq – A Python framework to work with high-throughput sequencing data. Bioinformatics 31, 166-169.

Antoniazza S, Burri R, Fumagalli L, Goudet J, Roulin A. (2010) Local adaptation maintains clinal variation in melanin-based coloration of European barn owls (Tyto alba). Evolution 64:1944-1954.

Ashburner M, Ball CA, Blake JA, et al. (2000) Gene Ontology: tool for the unification of biology. Nature Genetics 25, 25-29.

Axelsson E, Ratnakumar A, Arendt ML, Maqbool K, Webster MT, Perloski M, Liberg O, Arnemo JM, Hedhammar A, Lindblad-Toh K. (2013) The genomic signature of dog domestication reveals adaptation to a starch-rich diet. Nature 495:360-364.

Baldwin SA, Beal PR, Yao SY, King AE, Cass CE, Young JD. (2004) The equilibrative nucleoside transporter family, SLC29. Pflugers Archiv 447:735-743.

Barrett RD, Hoekstra HE. (2011) Molecular spandrels: tests of adaptation at the genetic level. Nataure Review Genetics 12:767-780.

Barrett RD, Schluter D (2008) Adaptation from standing genetic variation. Trends in Ecology & Evolution 23, 38-44.

Benjamini Y, Hochberg Y (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society. Series B (Methodological) 57, 289-300.

Bernatchez L, Dodson JJ. (1991) Phylogeographic Structure in Mitochondrial DNA of the Lake Whitefish (Coregonus Clupeaformis) and its relation to Pleistocene Glaciations. Evolution 45:1016-1035.

Berner D, Adams DC, Grandchamp AC, Hendry AP (2008) Natural selection drives patterns of lake-stream divergence in stickleback foraging morphology. Journal of Evolutionary Biology 21, 1653-1665.

Berner D, Roesti M, Hendry AP, Salzburger W (2010) Constraints on speciation suggested by comparing lake-stream stickleback divergence across two continents. Molecular Ecology 19, 4963-4978.

Bolnick DI, Snowberg LK, Caporaso JG, et al. (2014) Major Histocompatibility Complex class IIb polymorphism influences gut microbiota composition and diversity. Molecular Ecology 23, 4831-4845.

Brawand D, Wagner CE, Li YI, Malinsky M, Keller I, Fan S, Simakov O, Ng AY, Lim ZW, Bezault E, et al. (2014) The genomic substrate for adaptive radiation in African cichlid fish. Nature 513:375-381.

Brommer JE. (2011) Whither P_{ST}? The approximation of Qst by Pst in evolutionary and conservation biology. Journal of Evolutionary Biology 24:1160-1168.

Bryois J, Buil A, Evans DM, Kemp JP, Montgomery SB, Conrad DF, Ho KM, Ring S, Hurles M, Deloukas P, et al. (2014) Cis and trans effects of human genomic variants on gene expression. PLoS Genetics 10:e1004461.

Butlin RK, Saura M, Charrier G, et al. (2014) Parallel evolution of local adaptation and reproductive isolation in the face of gene flow. Evolution 68, 935-949.

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

Cardoso-Moreira M, Arguello JR, Gottipati S, Harshman LG, Grenier JK, Clark AG. (2016) Evidence for the fixation of gene duplications by positive selection in Drosophila. Genome Research 26:787-798.

Chain FJ, Feulner PG, Panchal M, et al. (2014) Extensive copy-number variation of young genes across stickleback populations. PLoS Genetics 10, e1004830.

Chan YF, Marks ME, Jones FC, et al. (2010) Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a Pitx1 enhancer. Science 327, 302-305.

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

Cheung VG, Spielman RS (2009) Genetics of human gene expression: mapping DNA variants that influence gene expression. Nature Reviews Genetics 10, 595-604.

Cheviron ZA, Whitehead A, Brumfield RT (2008) Transcriptomic variation and plasticity in rufous-collared sparrows (Zonotrichia capensis) along an altitudinal gradient. Molecular Ecology 17, 4556-4569.

Colombo M, Diepeveen ET, Muschick M, et al. (2013) The ecological and genetic basis of convergent thick-lipped phenotypes in cichlid fishes. Molecular Ecology 22, 670-684.

Colosimo PF, Hosemann KE, Balabhadra S, Villarreal G, Dickson M, Grimwood J, Schmutz J, Myers RM, Schluter D, Kingsley DM. (2005) Widespread Parallel Evolution in Sticklebacks by Repeated Fixation of Ectodysplasin Alleles. Science 307:1928-1933.

Conrad B, Antonarakis SE. (2007) Gene duplication: a drive for phenotypic diversity and cause of human disease. Annual Review of Genomics Human Genetics 8:17-35.

Conte GL, Arnegard ME, Peichel CL, Schluter D. (2012) The probability of genetic parallelism and convergence in natural populations. Proceedings of the Royal Society B: Biological Sciences.

Conti BJ, Davis BK, Zhang J, et al. (2005) CATERPILLER 16.2 (CLR16.2), a novel NBD/LRR family member that negatively regulates T cell function. The Journal of Biological Chemistry 280, 18375-18385.

Darwin C. (1859) On the Origin of Species.

Deagle BE, Jones FC, Chan YF, et al. (2012) Population genomics of parallel phenotypic evolution in stickleback across stream-lake ecological transitions. Proceedings of the Royal Society B: Biological Sciences 279, 1277-1286.

Derome N, Duchesne P, Bernatchez L (2006) Parallelism in gene transcription among sympatric lake whitefish (Coregonus clupeaformis Mitchill) ecotypes. Molecular Ecology 15, 1239-1249.

Dewar RC, Sherwin WB, Thomas E, Holleley CE, Nichols RA. (2011) Predictions of single-nucleotide polymorphism differentiation between two populations in terms of mutual information. Molecular Ecology 20:3156-3166.

Dillies MA, Rau A, Aubert J, et al. (2012) A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. Briefings in Bioinformatics 14, 671-683.

Divaris K, Monda KL, North KE, et al. (2012) Genome-wide association study of periodontal pathogen colonization. Journal of Dental Research 91, 21s-28s.

Dixon AL, Liang L, Moffatt MF, Chen W, Heath S, Wong KC, Taylor J, Burnett E, Gut I, Farrall M, et al. (2007) A genome-wide association study of global gene expression. Nature Genetics 39:1202-1207.

Donaldson-Matasci MC, Bergstrom CT, Lachmann M. (2010) The fitness value of information. Oikos 119:219-230.

Eizaguirre C, Lenz TL, Kalbe M, Milinski M (2012a) Rapid and adaptive evolution of MHC genes under parasite selection in experimental vertebrate populations. Nature Communications 3, 621.

Eizaguirre C, Lenz TL, Kalbe M, Milinski M (2012b) Divergent selection on locally adapted major histocompatibility complex immune genes experimentally proven in the field. Ecology Letters 15, 723-731.

Eizaguirre C, Lenz TL, Sommerfeld RD, et al. (2011) Parasite diversity, patterns of MHC II variation and olfactory based mate choice in diverging three-spined stickleback ecotypes. Evolutionary Ecology 25, 605-622.

Eizaguirre C, Lenz TL, Traulsen A, Milinski M (2009) Speciation accelerated and stabilized by pleiotropic major histocompatibility complex immunogenes. Ecology Letters 12, 5-12.

Elmer KR, Meyer A (2011) Adaptation in the age of ecological genomics: insights from parallelism and convergence. Trends in Ecology & Evolution 26, 298-306.

Eloy de Amorim M, Schoener TW, Santoro GRCC, Lins ACR, Piovia-Scott J, Brandão RA. (2017) Lizards on newly created islands independently and rapidly adapt in morphology and diet. Proceedings of the National Academy of Sciences 114:8812-8816.

Emerson JJ, Hsieh LC, Sung HM, Wang TY, Huang CJ, Lu HH, Lu MY, Wu SH, Li WH. (2010). Natural selection on cis and trans regulation in yeasts. Genome Research 20:826-836.

Enard W, Khaitovich P, Klose J, Zollner S, Heissig F, Giavalisco P, Nieselt-Struwe K, Muchmore E, Varki A, Ravid R, et al. (2002) Intra- and interspecific variation in primate gene expression patterns. Science 296:340-343.

Endler JA. (1986) Natural Selection in the Wild: Princeton University Press.

Excoffier L, Lischer HE. (2010) Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. Molecular Ecology Resources 10:564-567.

Feulner PG, Chain FJ, Panchal M, Eizaguirre C, Kalbe M, Lenz TL, Mundry M, Samonte IE, Stoll M, Milinski M, et al. (2013) Genome-wide patterns of standing genetic variation in a marine population of three-spined sticklebacks. Molecular Ecology 22:635-649.

Feulner PG, Chain FJ, Panchal M, et al. (2015) Genomics of Divergence along a Continuum of Parapatric Population Differentiation. PLoS Genetics 11, e1004966.

Ffrench-Constant RH, Rocheleau TA, Steichen JC, Chalmers AE (1993) A point mutation in a Drosophila GABA receptor confers insecticide resistance. Nature 363, 449-451.

Fischer R. (1930) The Genetical Theory of Natural Selection.

Flicek P, Amode MR, Barrell D, et al. (2012) Ensembl 2012. Nucleic Acids Research 40, D84-D90.

Frankel N, Erezyilmaz DF, McGregor AP, Wang S, Payre F, Stern DL. (2011) Morphological evolution caused by many subtle-effect substitutions in regulatory DNA. Nature 474:598-603.

Fraser HB, Moses AM, Schadt EE. (2010) Evidence for widespread adaptive evolution of gene expression in budding yeast. Proceedings of the National Academy of Sciences 107:2977-2982.

Franssen SU, Gu J, Bergmann N, et al. (2011) Transcriptomic resilience to global warming in the seagrass Zostera marina, a marine foundation species.

Proceedings of the National Academy of Sciences of the United States of America 108, 19276-19281.

Fraser HB (2013) Gene expression drives local adaptation in humans. Genome Research 23, 1089-1096.

Futuyma, DJ (1986) Evolutionary Biolgy. Sinaauer Associates.

Gamazon ER, Stranger BE. (2015) The impact of human copy number variation on gene expression. Briefings in Functional Genomics 14:352-357.

Gemayel R, Vinces MD, Legendre M, Verstrepen KJ (2010) Variable tandem repeats accelerate evolution of coding and regulatory sequences. Annual Review of Genetics 44, 445-477.

Gibson G (2008) The environmental contribution to gene expression profiles. Nature Reviews Genetics 9, 575-581.

Gibson G, Weir B (2005) The quantitative genetics of transcription. Trends in Genetics 21, 616-623.

Gilad Y, Rifkin SA, Pritchard JK (2008) Revealing the architecture of gene regulation: the promise of eQTL studies. Trends in Genetics 24, 408-415.

González-Prendes R, Quintanilla R, Cánovas A, Manunza A, Figueiredo Cardoso T, Jordana J, Noguera JL, Pena RN, Amills M. (2017) Joint QTL mapping and gene expression analysis identify positional candidate genes influencing pork quality traits. Scientific Reports 7:39830.

Haase D, Rieger JK, Witten A, et al. (2014) Specific gene expression responses to parasite genotypes reveal redundancy of innate immunity in vertebrates. PLoS One 9, e108001.

Haraksingh RR, Snyder MP (2013) Impacts of variation in the human genome on gene regulation. Journal of Molecular Biology 425, 3970-3977.

Hargeby A, Johansson J, Ahnesjo J (2004) Habitat-specific pigmentation in a freshwater isopod: adaptive evolution over a small spatiotemporal scale. Evolution 58, 81-94.

Harmon LJ, Kolbe JJ, Cheverud JM, Losos JB (2005) Convergence and the multidimensional niche. Evolution 59, 409-421.

Harrison PW, Wright AE, Mank JE (2012) The evolution of gene expression and the transcriptome-phenotype relationship. Seminars in Cell and Developmental Biology 23, 222-229.

Henrichsen CN, Chaignat E, Reymond A. (2009a) Copy number variants, diseases and gene expression. Human Molecular Genetics 18:R1-8.

Henrichsen CN, Vinckenbosch N, Zollner S, Chaignat E, Pradervand S, Schutz F, Ruedi M, Kaessmann H, Reymond A. (2009b) Segmental copy number variation shapes tissue transcriptomes. Nature Genetics 41:424-429.

Hill MO. (1973) Diversity and Evenness: A Unifying Notation and Its Consequences. Ecology 54:427-432.

Hirase S, Ozaki H, Iwasaki W. (2014) Parallel selection on gene copy number variations through evolution of three-spined stickleback genomes. BMC Genomics 15:735.

Hoekstra HE, Coyne JA. (2007) The locus of evolution: evo devo and the genetics of adaptation. 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, 101-104.

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:e1000862.

Houle D, Govindaraju DR, Omholt S. (2010) Phenomics: the next challenge. Nature Review Genetics 11:855-866.

Howe K, Clark MD, Torroja CF, Torrance J, Berthelot C, Muffato M, Collins JE, Humphray S, McLaren K, Matthews L, et al. (2013) The zebrafish reference genome sequence and its relationship to the human genome. Nature 496:498-503.

Hsing LC, Rudensky AY. (2005) The lysosomal cysteine proteases in MHC class II antigen presentation. Immunological Review 207:229-241.

Huddleston J, Eichler EE. (2016) An Incomplete Understanding of Human Genetic Variation. Genetics 202:1251-1254.

Huey RB, Gilchrist GW, Carlson ML, Berrigan D, Serra L. (2000) Rapid evolution of a geographic cline in size in an introduced fly. Science 287:308-309.

Ishikawa A, Kusakabe M, Yoshida K, Ravinet M, Makino T, Toyoda A, Fujiyama A, Kitano J. (2017) Different contributions of local- and distant-regulatory changes to transcriptome divergence between stickleback ecotypes. Evolution 71:565-581.

Iskow RC, Gokcumen O, Lee C. (2012) Exploring the role of copy number variants in human adaptation. Trends in Genetics 28:245-257.

Jones, R., Culver, D. C. and Kane, T. C. (1992) Are parallel morphologies of cave organisms the result of similar selection pressures? Evolution, 46: 353–365. doi:10.1111/j.1558-5646.1992.tb02043.x

Jones FC, Grabherr MG, Chan YF, Russell P, Mauceli E, Johnson J, Swofford R, Pirun M, Zody MC, White S, et al. (2012) The genomic basis of adaptive evolution in threespine sticklebacks. Nature 484:55-61.

Kaeuffer R, Peichel CL, Bolnick DI, Hendry AP (2012) Parallel and nonparallel aspects of ecological, phenotypic, and genetic divergence across replicate population pairs of lake and stream stickleback. Evolution 66, 402-418.

Kalbe M, Wegner KM, Reusch TBH (2002) Dispersion patterns of parasites in 0+ year three-spined sticklebacks: a cross population comparison. Journal of Fish Biology 60, 1529-1542.

Karvonen A, Lucek K, Marques DA, Seehausen O (2015) Divergent macroparasite infections in parapatric Swiss lake-stream pairs of threespine stickleback (*Gasterosteus aculeatus*). PLoS One 10, e0130579.

Kedzierski L, Montgomery J, Curtis J, Handman E (2004) Leucine-rich repeats in host-pathogen interactions. Archivum Immunologiae et Therapia Experimentalis 52, 104-112.

Khaitovich P, Paabo S, Weiss G. (2005) Toward a neutral evolutionary model of gene expression. Genetics 170:929-939.

Kim D, Pertea G, Trapnell C, et al. (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biology 14, R36.

Kim T, Kim K, Lee SH, et al. (2009) Identification of LRRc17 as a Negative Regulator of Receptor Activator of NF- κ B Ligand (RANKL)-induced Osteoclast Differentiation. Journal of Biological Chemistry 284, 15308-15316.

Kimura M. (1968) Evolutionary rate at the molecular level. Nature 217, 624–626

Kimura M. (1983) The Neutral Theory of Molecular Evolution. Cambridge University Press, Cambridge.

King MC, Wilson AC. (1975) Evolution at two levels in humans and chimpanzees. Science 188:107-116.

Kreitman, M. (1996), The neutral theory is dead. Long live the neutral theory. Bioessays, 18: 678–683. doi:10.1002/bies.950180812

Krücken J, Schroetel RMU, Müller IU, Saïdani N, Marinovski P, Benten WPM, Stamm O, Wunderlich F. (2004) Comparative analysis of the human gimap gene cluster encoding a novel GTPase family. Gene 341:291-304.

Laland K, Uller T, Feldman M, Sterelny K, Muller GB, Moczek A, Jablonka E, Odling-Smee J, Wray GA, Hoekstra HE, et al. (2014) Does evolutionary theory need a rethink? Nature 514:161-164.

Leder EH, McCairns RJ, Leinonen T, et al. (2014) The Evolution and Adaptive Potential of Transcriptional Variation in Sticklebacks-Signatures of Selection and Widespread Heritability. Molecular Biology and Evolution 32, 674-689.

Leinonen T, McCairns RJ, O'Hara RB, Merila J. (2013) QST-FST comparisons: evolutionary and ecological insights from genomic heterogeneity. Nature Review Genetics 14:179-190.

Lemos B, Meiklejohn CD, Caceres M, Hartl DL. (2005) Rates of divergence in gene expression profiles of primates, mice, and flies: stabilizing selection and variability among functional categories. Evolution 59:126-137.

Lenz TL (2015) Transcription in space - environmental vs. genetic effects on differential immune gene expression. Molecular Ecology 24, 4583-4585.

Lenz TL, Eizaguirre C, Rotter B, Kalbe M, Milinski M (2013) Exploring local immunological adaptation of two stickleback ecotypes by experimental infection and transcriptome-wide digital gene expression analysis. Molecular Ecology 22, 774-786.

Lewontin RC, Krakauer J. (1973) Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. Genetics 74:175-195.

Linnen CR, Kingsley EP, Jensen JD, Hoekstra HE. (2009) On the origin and spread of an adaptive allele in deer mice. Science 325:1095-1098.

Losos JB, Jackman TR, Larson A, Queiroz K, Rodriguez-Schettino L (1998) Contingency and determinism in replicated adaptive radiations of island lizards. Science 279, 2115-2118.

Love MI, Huber W, Anders S (2014) Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2. Genome Biology 15, 550

Lucek K, Lemoine M, Seehausen O (2014) Contemporary ecotypic divergence during a recent range expansion was facilitated by adaptive introgression. Journal of Evolutionary Biology 27, 2233-2248.

Machado LR, Ottolini B. (2015) An evolutionary history of defensins: a role for copy number variation in maximizing host innate and adaptive immune responses. Frontiers in Immunology 6:115.

Machado-Schiaffino G, Henning F, Meyer A (2014) Species-specific differences in adaptive phenotypic plasticity in an ecologically relevant trophic trait: hypertrophic lips in Midas cichlid fishes. Evolution 68, 2086-2091.

MacKinnon JS, Rundle HD (2002) Speciation in nature: the threespine stickleback model systems. Trends in Ecology and Evolution 17, 480-488.

Mallarino R, Linden TA, Linnen CR, Hoekstra HE. (2017) The role of isoforms in the evolution of cryptic coloration in Peromyscus mice. Molecular Ecology 26:245-258.

Manceau M, Domingues VS, Mallarino R, Hoekstra HE (2011) The developmental role of Agouti in color pattern evolution. Science 331, 1062-1065.

Manousaki T, Hull PM, Kusche H, et al. (2013) Parsing parallel evolution: ecological divergence and differential gene expression in the adaptive radiations of thick-lipped Midas cichlid fishes from Nicaragua. Molecular Ecology 22, 650-669.

Merilä J, Crnokrak P. (2001) Comparison of genetic differentiation at marker loci and quantitative traits. Journal of Evolutionary Biology 14:892-903.

Metzger BPH, Wittkopp PJ, Coolon JD. (2017) Evolutionary Dynamics of Regulatory Changes Underlying Gene Expression Divergence among Saccharomyces Species. Genome Biology and Evolution 9:843-854.

Morris MR, Richard R, Leder EH, et al. (2014) Gene expression plasticity evolves in response to colonization of freshwater lakes in threespine stickleback. Molecular Ecology 23, 3226-3240.

Moser D, Kueng B, Berner D (2015) Lake-stream divergence in stickleback life history: a plastic response to trophic niche differentiation? Evolutionary Biology 42, 328-338.

Muschick M, Indermaur A, Salzburger W (2012) Convergent evolution within an adaptive radiation of cichlid fishes. Current Biology 22, 2362-2368.

Nosil P, Feder JL. (2012). Genomic divergence during speciation: causes and consequences. Philos Trans R Soc Lond B Biological Sciences367:332-342.

Nourmohammad A, Rambeau J, Held T, Kovacova V, Berg J, Lässig M. Adaptive evolution of gene expression in Drosophila. Cell Reports 20:1385-1395.

Orr, H. (1998). The population genetics of adaptation: The distribution of factors fixed during adaptive evolution. Evolution, 52(4), 935-949. doi:10.2307/2411226

Oleksiak MF, Churchill GA, Crawford DL. 2002. Variation in gene expression within and among natural populations. Nature Genetics 32:261-266.

Pavey SA, Collin H, Nosil P, Rogers SM (2010) The role of gene expression in ecological speciation. Annals of the New York Academy of Sciences 1206, 110-129.

Pavey SA, Sutherland BJ, Leong J, et al. (2011) Ecological transcriptomics of lake-type and riverine sockeye salmon (Oncorhynchus nerka). BMC Ecology 11, 31.

Perry GH, Dominy NJ, Claw KG, Lee AS, Fiegler H, Redon R, Werner J, Villanea FA, Mountain JL, Misra R, et al. (2007) Diet and the evolution of human amylase gene copy number variation. Nature Genetics 39:1256-1260.

Pigeon D, Chouinard A, Bernatchez L. (1997) Multiple modes of speciation involved in the parallel evolution of sympatric morphotypes of lake whitefish (coregonus clupeaformis, salmonidae). Evolution 51:196-205.

Prabhakar S, Visel A, Akiyama JA, Shoukry M, Lewis KD, Holt A, Plajzer-Frick I, Morrison H, Fitzpatrick DR, Afzal V, et al. (2008) Human-specific gain of function in a developmental enhancer. Science 321:1346-1350.

Press C, Evensen O (1999) The morphology of the immune system in teleost fishes. Fish & Shellfish Immunology 9, 309-318.

Pritchard JK, Pickrell JK, Coop G (2010) The Genetics of Human Adaptation: Hard Sweeps, Soft Sweeps, and Polygenic Adaptation. Current Biology 20, R208-R215.

Pritchard VL, Viitaniemi HM, McCairns RJ, Merila J, Nikinmaa M, Primmer CR, Leder EH. (2017) Regulatory Architecture of Gene Expression Variation in the Threespine Stickleback Gasterosteus aculeatus. G3 (Bethesda) 7:165-178.

Protas ME, Hersey C, Kochanek D, et al. (2006) Genetic analysis of cavefish reveals molecular convergence in the evolution of albinism. Nature Genetics 38, 107-111.

Pujolar JM, Ferchaud AL, Bekkevold D, Hansen MM. (2017) Non-parallel divergence across freshwater and marine three-spined stickleback Gasterosteus aculeatus populations. Journal of Fish Biology 91:175-194.

R Development Core Team (2008) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org

Ranz JM, Machado CA (2006) Uncovering evolutionary patterns of gene expression using microarrays. Trends in Ecology and Evolution 21, 29-37.

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, 1663-1667.

Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH, Carson AR, Chen W, et al. (2006) Global variation in copy number in the human genome. Nature 444:444-454.

Renaut S, Maillet N, Normandeau E, Sauvage C, Derome N, Rogers SM, Bernatchez L. (2012) Genome-wide patterns of divergence during speciation: the lake whitefish case study. Philosophical Transactions of the Royal Society London B Biological Sciences 367:354-363.

Reusch TB, Wegner KM, Kalbe M (2001) Rapid genetic divergence in postglacial populations of threespine stickleback (Gasterosteus aculeatus): the role of habitat type, drainage and geographical proximity. Molecular Ecology 10, 2435-2445.

Rice AM, McLysaght A. (2017) Dosage sensitivity is a major determinant of human copy number variant pathogenicity. Nature Communications 8:14366.

Rifkin SA, Kim J, White KP. (2003) Evolution of gene expression in the Drosophila melanogaster subgroup. Nature Genetics 33:138-144.

Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26, 139-140.

Robinson MD, Oshlack A (2010) A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biology 11, R25.

Roesti M, Salzburger W, Berner D. (2012a) Uninformative polymorphisms bias genome scans for signatures of selection. BMC Evolutionary Biology 12:94.

Roesti M, Hendry AP, Salzburger W, Berner D (2012b) Genome divergence during evolutionary diversification as revealed in replicate lake-stream stickleback population pairs. Molecular Ecology 21, 2852-2862.

Rohlenová K, Morand S, Hyršl P, et al. (2011) Are fish immune systems really affected by parasites? an immunoecological study of common carp (Cyprinus carpio). Parasites & Vectors 4, 120.

Savolainen O, Lascoux M, Merila J (2013) Ecological genomics of local adaptation. Nature Reviews Genetics 14, 807-820.

Scharsack JP, Kalbe M, Harrod C, Rauch G (2007) Habitat-specific adaptation of immune responses of stickleback (Gasterosteus aculeatus) lake and river ecotypes. Proceedings of the Royal Society B: Biological Sciences 274, 1523-1532.

Schena M, Shalon D, Davis RW, Brown PO. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 270:467-470.

Schlattl A, Anders S, Waszak SM, Huber W, Korbel JO. (2011) Relating CNVs to transcriptome data at fine resolution: assessment of the effect of variant size, type, and overlap with functional regions. Genome Research 21:2004-2013.

Schluter D, Clifford EA, Nemethy M, McKinnon JS. (2004) Parallel evolution and inheritance of quantitative traits. American Naturalist 163:809-822.

Schmieder R, Edwards R (2011) Quality control and preprocessing of metagenomic datasets. Bioinformatics 27, 863-864.

Schneider M, Zimmermann AG, Roberts RA, et al. (2012) The innate immune sensor NLRC3 attenuates Toll-like receptor signaling via modification of the signaling

adaptor TRAF6 and transcription factor NF-kappaB. Nature Immunology 13, 823-831.

Schrider DR, Hahn MW, Begun DJ. (2016) Parallel Evolution of Copy-Number Variation across Continents in Drosophila melanogaster. Molecular Biology and Evolution 33:1308-1316.

Schwefel D, Fröhlich C, Eichhorst J, Wiesner B, Behlke J, Aravind L, Daumke O. (2010) Structural basis of oligomerization in septin-like GTPase of immunity-associated protein 2 (GIMAP2). Proceedings of the National Academy of Sciences of the United States of America 107:20299-20304.

Shen Y, Yue F, McCleary DF, Ye Z, Edsall L, Kuan S, Wagner U, Dixon J, Lee L, Lobanenkov VV, et al. (2012) A map of the cis-regulatory sequences in the mouse genome. Nature 488:116-120.

Slatkin M, Voelm L. (1991) F_{ST} in a hierarchical island model. Genetics 127:627-629.

Stamatoyannopoulos JA (2004) The genomics of gene expression. Genomics 84, 449-457.

Stanley ER, Cifone M, Heard PM, Defendi V (1976) Factors regulating macrophage production and growth: identity of colony-stimulating factor and macrophage growth factor. Journal of Expeirmental Medicine 143, 631-647.

Stern DL. (2013) The genetic causes of convergent evolution. Nature Review Genetics 14:751-764.

Stranger BE, Nica AC, Forrest MS, Dimas A, Bird CP, Beazley C, Ingle CE, Dunning M, Flicek P, Koller D, et al. (2007a) Population genomics of human gene expression. Nature Genetics 39:1217-1224.

Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, Thorne N, Redon R, Bird CP, de Grassi A, Lee C, et al. (2007b) Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science 315:848-853.

Stuart YE, Veen T, Weber JN, Hanson D, Ravinet M, Lohman BK, Thompson CJ, Tasneem T, Doggett A, Izen R, et al. (2017) Contrasting effects of environment and genetics generate a continuum of parallel evolution. Nature Ecology and Evolution 1:158.

Stutz WE, Schmerer M, Coates JL, Bolnick DI (2015) Among-lake reciprocal transplants induce convergent expression of immune genes in threespine stickleback. Molecular Ecology.

Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G, Hsi-Yang Fritz M, et al. (2015) An integrated map of structural variation in 2,504 human genomes. Nature 526:75-81.

Timur ZK, Akyildiz Demir S, Seyrantepe V. (2016) Lysosomal Cathepsin A Plays a Significant Role in the Processing of Endogenous Bioactive Peptides. Frontiers Molecular Biosciences 3:68.

Turesson, G. (1922) The species and the variety as ecological units. Hereditas, 3: 100–113. doi:10.1111/j.1601-5223.1922.tb02727.x

Turrill, W. B. (1946) The ecotype concept. New Phytologist, 45: 34–43. doi:10.1111/j.1469-8137.1946.tb05044.x

Uebbing S, Kunstner A, Makinen H, Backstrom N, Bolivar P, Burri R, Dutoit L, Mugal CF, Nater A, Aken B, et al. (2016) Divergence in gene expression within and between two closely related flycatcher species. Molecular Ecology 25:2015-2028.

Veitia RA. (2002) Exploring the etiology of haploinsufficiency. Bioessays 24:175-184.

Veitia RA. (2005) Gene dosage balance: deletions, duplications and dominance. Trends Genetics 21:33-35.

Wegner KM, Reusch TB, Kalbe M (2003) Multiple parasites are driving major histocompatibility complex polymorphism in the wild. Journal of Evolutionary Biology 16, 224-232.

Whitehead A (2012) Comparative genomics in ecological physiology: toward a more nuanced understanding of acclimation and adaptation. The Journal of Experimental Biology 215, 884-891.

Whitehead A, Crawford DL. (2006) Neutral and adaptive variation in gene expression. Proceedings of the National Academy of Sciences 103:5425-5430.

Wittkopp PJ, Haerum BK, Clark AG. (2008) Regulatory changes underlying expression differences within and between Drosophila species. Nature Genetics 40:346-350.

Wittkopp PJ, Kalay G. (2011) Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. Nature Review Genetics 13:59-69.

Wood, T. E., Burke, J. M., & Rieseberg, L. H. (2005). Parallel genotypic adaptation: when evolution repeats itself. Genetica, 123(1-2), 157–170.

Wray GA. (2007) The evolutionary significance of cis-regulatory mutations. Nature Review Genetics 8:206-216.

Wright S. (1951) The genetical structure of populations. Ann Eugen 15:323-354.

Zhang L, Mo J, Swanson KV, et al. (2014) NLRC3, a member of the NLR family of proteins, is a negative regulator of innate immune signaling induced by the DNA sensor STING. Immunity 40, 329-341.

Zhao L, Wit J, Svetec N, Begun DJ. (2015) Parallel Gene Expression Differences between Low and High Latitude Populations of Drosophila melanogaster and D. simulans. PLoS Genetics 11:e1005184.

Acknowledgements

I thank Prof. Manfred Milinski for offering me this opportunity to pursuit a doctoral degree in the field of Evolutionary Biology. When I first joined the group, I was completely new to the field. Through these years, I have learnt a lot from Prof. Milinski's vast knowledge and got to know so many diverse and interesting topics in Evolutionary Biology.

I genuinely thank my direct supervisors, Frederic Chain and Philine Feulner, for their continuing commitment and efforts into supervising my PhD work. Thanks to their unequivocal and intense supervision, I have grown from a naïve student to a qualified candidate. They themselves set good examples of diligence, efficiency and rigorous scholarship for me, which also surely benefit my future career. I also thank them for their unlimited patience guiding me through the difficult time. Having both of them on board for supervising my PhD study is a big luck for me.

I thank all the co-authors I collaborated with. These include Mahesh Panchal, Irene Samonte, Christophe Eizaguirre, Tobias Lenz, Martin Kalbe, and the senior researchers, Prof. Monika Stoll, Prof. Erich Bornberg-Bauer and Prof. Thorsten Reusch. Thanks for the discussions, suggestions and feedbacks, which greatly help developing ideas and structuring the manuscripts.

I thank all the colleagues working in the institute for making it such an open ambiance to do research. I thank the department members, e.g. Joshka Kaufmann, Noemie Erin, Nina Hafer, for discussions at seminars. I thank Agnes Piecyk for the mutual encouragements when sharing the office. I thank Dominik Schmid for the help in translating the summary of this thesis into German. I thank Britta Barron and Petra Salenz and others for the helps in administrative fairs. I thank Derk Wachsmuth for taking care of the computing server, which I heavily rely on. I also thank International Max Planck Research School, and especially Kerstin Mehnert for her great help.

My time in Plön not only witnesses my academic pursuit, but also varieties of culture exchange activities. To name a few, PubQuiz, Halloween parties, movie nights, Friday BBQs, and all kinds of sports. I enjoy these MPI traditions a lot, thanks to the voluntary organizers. Besides, I also thank my Chinese friends in Plön, whom I felt like family when living abroad. These include Bin Wu, Jie Cheng, Weini Huang, Chen Xie, and so on.

My gratitude also goes to my friends out of the academic world, e.g., my travel mate and foodie friends. Thanks for all the adventures.

I am deeply in debt to my parents, who provide me unconditional and unreserved support and love through my life.

Last but not least, special thanks to my husband, Pin-Jui Hsu, for being there with me through the ups and downs during the second half of my PhD time.

Curriculum Vitæ

Yun Huang

July 02, 1986	Born in Fuzhou, China Nationality: Chinese
2004 – 2008	Bachelor of Science: Pharmacy Soochow University, China
2009 – 2012	Master of Science: Life Science Informatics University of Bonn, Germany Master Thesis: Identification and Characterization of Activity Ridges in Large Data Sets Supervisor: Prof. Dr. Jürgen Bajorath
2012 – 2017	PhD candidate Max Planck Institute for Evolutionary Biology, Plön Supervisor: Prof. Dr. Manfred Milinski

Declaration

Hereby I declare that

1) apart from my supervisor's guidance, the content and design of this dissertation is the product of my own work. The co-authors' contributions to specific chapters are listed in the thesis outline section.
2) this thesis has not been submitted either partially or wholly as part of a doctoral degree to another examination body, and no other materials are published or submitted for publication than indicated in the thesis.
3) the preparation of the thesis has been subjected to the Rules of Good Scientific Practice of the German Research Foundation.

Plön 8th of November 2017

Yun Huang