Genome-wide genotype-expression relationships reveal both copy number and single nucleotide differentiation contribute to differential gene expression between stickleback ecotypes

Yun Huang[1,2*], Philine GD Feulner[3,4], Christophe Eizaguirre[5], Tobias L Lenz[1], Erich Bornberg-Bauer[6], Manfred Milinski[1], Thorsten BH Reusch[7], Frédéric JJ Chain[8*]

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

[2] Biodiversity Research Center, Academia Sinica, Taipei, Taiwan, ROC

[3] 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

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

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

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

[7] Marine Evolutionary Ecology, GEOMAR Helmholtz Centre for Ocean Research Kiel, 24105 Kiel, Germany

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

[*] Correspondence to: yunhuang@gate.sinica.edu.tw ; Frederic_Chain@uml.edu

Keywords

habitat-specific adaptation, CNV, copy number variation, eSNP, *cis*-regulatory regions, expression differentiation, three-spined stickleback

Abstract

Repeated and independent emergence of trait divergence that matches habitat differences is a sign of parallel evolution by natural selection. Yet, the molecular underpinnings that are targeted by adaptive evolution often remain elusive. We investigate this question by combining genome-wide analyses of copy number variants (CNVs), single nucleotide polymorphisms (SNPs), and gene expression across four pairs of lake and river populations of the three-spined stickleback (*Gasterosteus aculeatus*). We tested whether CNVs that span entire genes and SNPs occurring in putative *cis*-regulatory regions contribute to gene expression differences between sticklebacks from lake and river origins. We found 135 gene CNVs that showed a significant positive association between gene copy number and gene expression, suggesting that CNVs result in dosage effects that can fuel phenotypic variation and serve as substrates for habitat-specific selection. Copy number differentiation between lake and river sticklebacks also contributed to expression differences of two immune-related genes in immune tissues, *cathepsin A* and *GIMAP7*. In addition, we identified SNPs in *cis*-regulatory regions (eSNPs) associated with the expression of 1865 genes, including one eSNP upstream of a carboxypeptidase gene where both the SNP alleles differentiated and the gene was differentially expressed between lake and river populations. Our study highlights two types of mutations as important sources of genetic variation involved in the evolution of

gene expression and in potentially facilitating repeated adaptation to novel environments.

Introduction

Uncovering the genetic mechanisms underlying adaptive evolution is a major research focus in evolutionary biology (Barrett and Hoekstra 2011). Adaptive phenotypes can result from changes in amino acid sequences that affect protein structure and function (Hoekstra and Coyne 2007), as well as from alterations of gene expression patterns (Carroll 2008). While gene expression can be plastic and respond to environmental stimuli (Gibson 2008), adaptive evolution of gene expression rests upon an inherited genetic basis.

Gene expression differences between populations and species often carry a significant heritable component and impact fitness, contributing to adaptation (Stamatoyannopoulos 2004; Whitehead and Crawford 2006b; Pavey, et al. 2010). A growing body of evidence has linked the acquisition of adaptive phenotypes in new environments 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), convergent thick lips in cichlids (Colombo, et al. 2013), and repeated pelvic loss in three-spined sticklebacks (Chan, et al. 2010). If the differentiation in expression confers an adaptive advantage across independent population clines, it may lead to parallel evolution at the gene expression level. The parallel evolution of expression patterns can be directly inferred when, for example, heritable gene expression variation correlates with an environmental cline rather than by ancestry (Whitehead and Crawford 2006a; Lenz 2015). Parallel gene expression has been observed in a few cases of diverging

ecotypes or species of adaptive radiations (Derome, et al. 2006; Pavey, et al. 2011; Colombo, et al. 2013; Manousaki, et al. 2013; Stutz, et al. 2015; Zhao, et al. 2015; Hanson, et al. 2017). Yet, the genetic variants associated with these gene expression patterns remain understudied.

Genomic studies of recurring ecotypes have revealed a major contribution of regulatory regions to parallel genomic divergence (Jones, et al. 2012; Brawand, et al. 2014). Combining gene expression surveys with genome-wide sequence analysis allows evaluating the role of genetic variants on the evolution of expression differences between ecotypes. The genetic basis of expression differences may reside in close physical proximity of a gene (in *cis*) or far away (in *trans*) (Gilad, et al. 2008). Genetic mutations altering the sequence of *cis*-regulatory elements can affect the binding affinity of transcription factors, whose effects are mainly limited to expression variation levels of neighboring genes, whereas mutations that affect *trans*-regulatory elements typically encode transcription factors that regulate multiple downstream genes (Wittkopp and Kalay 2011). Due to the local effects of *cis*-regulatory elements that confer a lower extent of pleiotropy compared to *trans*-, *cis*-regulatory elements have been suggested to be more important than *trans*-regulatory elements in the expression divergence between species (Wittkopp, et al. 2008).

In addition to sequence changes in its regulatory region, the number of copies of a particular gene can affect its expression. Gene copy number can differ among individuals due to genetic deletions and duplications, giving rise to copy number variations (CNVs), which natural selection can act upon (Nguyen, et al. 2006; Katju and Bergthorsson 2013). Copy number is generally positively correlated with expression levels (Haraksingh and Snyder 2013; Gamazon and Stranger 2015), producing a gene dosage effect (Zhang, et al. 2009). Gene dosage effects are often detrimental to fitness as they can disrupt the

stoichiometric balance in molecular networks (Papp, et al. 2003; Veitia 2005; Birchler and Veitia 2012) and have been associated with diseases (Rice and McLysaght 2017). However, in some cases, dosage effects of CNVs have also been beneficial, such as the relationship observed between amylase gene copy numbers and starch diets in both humans and dogs (Perry, et al. 2007; Axelsson, et al. 2013), and the number of cytochrome P450 genes in insecticide-resistant populations of dengue mosquitos (Faucon, et al. 2015). While variation in *cis*-regulatory elements and CNVs can both affect gene expression and contribute to adaptive phenotypes, their contribution to habitat-specific gene expression has not been systematically studied. Genotype-expression relationships become particularly interesting when divergence patterns across replicated populations independently adapted to different environments occur at both the genetic and expression levels, strongly suggesting a genetic basis underlying adaptive expression variation.

The three-spined stickleback (*Gasterosteus aculeatus*) is a powerful model species to investigate habitat-specific adaptation. After the last glaciation, marine three-spined sticklebacks repeatedly colonized different freshwater habitats, resulting in an adaptive radiation composed of habitat-specific ecotypes (McKinnon and Rundle 2002). In particular, recurrent adaptation to lakes and rivers (or streams) has given rise to distinct ecotypes across the northern hemisphere (Reusch, et al. 2001), with morphological differences in body shapes and traits involved in foraging (Berner, et al. 2008; Deagle, et al. 2012; Kaeuffer, et al. 2012; Ravinet, et al. 2013; Lucek, et al. 2014). Another profound difference between lake and river habitats is the distinct parasite communities, in which lake fish generally suffer from a higher parasite burden than river fish, likely contributing to recurrent ecotype differences at both the phenotypic and genetic level (Kalbe, et al. 2002; Eizaguirre, et al. 2011; Feulner, et al. 2015). Transcriptome analyses have revealed

over a hundred genes with habitat-specific gene expression among wild-caught lake and river sticklebacks (Huang, et al. 2016), some of which were also differentially expressed between lake and river sticklebacks in a laboratory-controlled parasite infection experiment (Lenz, et al. 2013). Lake and stream sticklebacks raised in common garden conditions also exhibit parallel gene expression differences (Hanson, et al. 2017). This suggests a heritable component to habitat-specific gene expression, which is also supported by quantitative genetics analyses on pedigrees of sticklebacks (Leder, et al. 2015). In sticklebacks, a greater contribution of *cis*-regulatory elements than *trans*-regulatory elements in expression variation and divergence between ecotypes has been suggested (Ishikawa, et al. 2017; Pritchard, et al. 2017; Verta and Jones 2019). However, unlike the parallel divergence observed between marine and freshwater sticklebacks at the sequence level (Jones et al. 2012) and in gene CNVs (Hirase et al. 2014), a low degree of parallel genetic differentiation exists among repeatedly diverged lake and river ecotypes, both at the sequence level (Deagle, et al. 2012; Roesti, et al. 2012; Feulner, et al. 2015; Stuart, et al. 2017) and in copy numbers (Chain, et al. 2014). This is despite habitat-specific patterns of gene expression (Huang, et al. 2016; Hanson, et al. 2017). Given the low degree of genomic parallelism, the genetic variation underlying the expression divergence between lake and river ecotypes remains elusive.

In this study, stickleback genomes and transcriptomes from the exact same individuals were used to study the molecular basis of habitat-specific adaptations between lake and river ecotypes. To identify candidate genes involved in adaptation to distinct parasite communities in lakes and rivers, we evaluated the relationships between gene expression variation in immune tissues and two types of variants, gene CNVs and SNPs in *cis*-regulatory regions. We tested for (1) associations between gene copy numbers or SNP genotypes and gene expression within and across individuals, (2) evidence of

habitat-specific selection as inferred from different gene copy numbers between ecotypes or allele frequency differentiation of SNPs, and (3) differential gene expression between ecotypes. These serve as three pillars of evidence that genetic changes contribute to adaptive gene expression differences between ecotypes. In these ways, we identified genetic variants that influence repeated differential expression between ecotypes, putatively contributing to habitat-specific adaptation.

Methods

Sampling design

To study the genetic differentiation between lake and river stickleback ecotypes that underlie expression differentiation, we combined a whole genome and a whole transcriptome dataset from a total of eight geographically widespread populations of three-spined sticklebacks that had been previously analyzed separately. The whole-genome sequence dataset consisted of 48 fish 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 each river (_R), respectively (Feulner, et al. 2015, Chain, et al. 2014, EBI Accession no: PRJEB5198; Figure S1; Table S1). The average genomic coverage was 26-fold, and genotypes from the whole-genome sequencing were validated with 98% concordance by Illumina's Golden Gate assay (Feulner, et al. 2015), yielding reliable SNP data reused in this study. The whole-transcriptome dataset comprised gene expression data from a subset of the same individuals as referenced above (43 total fish, matched IDs indicated in Table S1). These transcriptomes were previously used to investigate habitat-specific gene expression between lake and river ecotypes (Huang, et al. 2016). To understand the adaptation to distinct parasite environments between lake and river habitats, we focused on two

immune tissues: we used 40 head kidney transcriptomes and 36 spleen transcriptomes (PRJEB8677). The average transcriptome library size was 6.5 million pair-end reads, which has limited power to detect genes with low expression but should be robust to quantify differences among medium to highly expressed genes (Tarazona et al. 2011; Ching et al. 2014).

Expression profiling

Transcriptome libraries from the sampled populations were first analyzed following (Huang, et al. 2016; Dryad doi:10.5061/dryad.hq50s). In short, transcriptome libraries from head kidneys and from spleens were analyzed separately. Weakly expressed genes with less than one read count per million in at least half of the respective tissue samples were removed and then libraries were normalized using the trimmed mean of M-value (TMM) method (Robinson and Oshlack 2010) in EdgeR (Robinson, et al. 2010). Expression levels were estimated as the log of normalized read count per million. The final set of expression profiles consisted of 12,105 genes from the head kidney and 12,451 genes from the spleen that were used in the analyses described below.

Identification of gene eCNVs

CNV regions of the study populations were identified by Chain *et al.* (2014), where CNVs were assigned using consensus calls from the read depth approach implemented in the software CNVnator (Abyzov, et al. 2011) and at least one other approach (paired-end and split-reads; for details see Chain et al. 2014). We identified genes with at least 95% length overlap with CNVs. Gene copy number was estimated using CNVnator and rounded to the closest integer. Genes showing no variation in estimated copy numbers amongst individuals of our study were excluded from copy number analyses. Genes with copy

number estimates of zero but with detectable read depth above zero were removed from our analyses to avoid possible false deletion calls. A total of 832 autosomal protein-coding genes remained, referred herein as "gene CNVs".

Using gene copy numbers and the corresponding gene expression from the same fish, we evaluated the association between gene copy number and expression level for each gene CNV in each individual, and for each tissue type separately. Using a linear mixed effect model, gene copy number was set as a fixed effect, and the population and sex were set as random effects (expression levels ~copy_number + (1|population) + (1|sex)). This approach makes use of the continuous nature of copy number genotypes and tests for dosage effects of CNVs, which is different from the typical eQTL approach that associates expression variation to categorical genotypes. Benjamini-Hochberg's multiple test correction was applied to the p-values of the fixed effect of copy number (Benjamini and Hochberg 1995). Genes with corrected p-values smaller than 0.05 were considered as "gene eCNVs", having statistically significant correlations between copy number and expression.

Identification of eSNPs

In addition to the evaluation of gene eCNVs, we mapped SNPs in *cis*-regulatory regions (eSNPs) to identify potential *cis*-regulatory elements that underlie gene expression variation. The eSNPs were determined for gene expression in head kidney and spleen separately, using SNPs within a 5kb range of the transcription start sites (TSSs). We reasoned that the 5kb upstream regions serve as a proxy for the location of potential *cis*-regulatory elements, based on empirical findings of cis-regulatory sequences in mouse (Shen, et al. 2012). SNPs used in this study were extracted from a previous genome-wide survey (Feulner, et al. 2015), excluding SNPs in CNV regions due to potential detection

biases (Hartasánchez et al. 2018) and filtering SNPs for a minor allele frequency greater than 0.05 in the four population pairs combined. Out of 12,105 and 12,451 genes expressed in the head kidney and in the spleen respectively, 10,803 and 10,914 genes had a total of 815,341 and 841,063 SNPs, and jointly 870,917 SNPs that fulfilled our filtering criteria. For each expressed gene, we tested for a significant association between each SNP and expression levels in FastQTL v2.165 (Ongen, et al. 2016) using the nominal pass and correcting for population stratification (population pairs and habitats) and sex. Two steps of multiple testing correction (Benjamini-Hochberg) were applied on the p-value for each SNP: the p-values were first corrected for numbers of SNPs per gene and then for the total number of genes tested. SNPs with corrected p-values smaller than 0.05 were considered as eSNPs.

Expression differentiation between stickleback ecotypes

Differential expression (DE) analyses implemented in the package EdgeR was previously used to identify significantly differentially expressed genes between ecotypes, indicative of habitat-specific gene expression (Huang, et al. 2016). To complement this binary categorization, we quantified the extent of expression differentiation in a continuous manner by computing the variable P_{CT}, which evaluates the relative variance in expression between groups compared to the variance within groups. We calculated P_{CT} between lake and river sticklebacks and accounted for expression variances among geographic population pairs and between sex using an ANOVA-based approach (methods adapted from Uebbing, et al. 2016). P_{CT} as a measure of relative differentiation in gene expression between lake and river ecotypes was calculated for each expressed gene and for the head kidney and spleen separately. Because the calculation of P_{CT} is conceptually equivalent to the calculation of copy number differentiation (V_{CT}, see below) and

nucleotide differentiation (F_{CT}, see below), the evaluation of expression differentiation is made directly comparable to that of genetic differentiation. We applied 1000 permutations following the methods for V_{CT} to identify genes with significant P_{CT}, and p-values were corrected by the Benjamini-Hochberg method (Benjamini and Hochberg 1995) for numbers of genes tested. Genes with adjusted p-values smaller than 0.05 for P_{CT} were considered significant. For candidate genes, we also calculated P_{CT} between each population pair, in the same way that P_{CT} was calculated for all populations combined but without population structure in the ANOVA model.

Copy number differentiation of gene eCNVs

In order to investigate the contribution of gene eCNVs in expression differentiation, we evaluated copy number differentiation between ecotypes across all population pairs together. For each gene eCNV, we calculated V_{CT} representing the relative variance in copy number between groups (here lake versus river ecotypes) compared to the overall variance within groups, similarly to P_{CT}. V_{CT} was calculated using all individuals from the 4 population pairs with an ANOVA-based approach, where lake and river ecotypes were treated as two comparison groups, while accounting for variance between population pairs (copy_number ~ ecotypes * population_pair). As we exclude CNVs in the sex chromosome for our analyses, we did not include sex as a factor in the model. V_{CT} is different from V_{ST}, a measurement of copy number differentiation between populations without a nested structure (Redon, et al. 2006), which was previously calculated on the same data set but between each lake and river pair separately in Chain et al. (2014). Including all population pairs together to estimate copy number differentiation (V_{CT}) detects overall increases or decreases of copy number across all pairs and increases sensitivity to detect such patterns, as it does not require differentiation

signals to be extreme in each pair. To determine how likely each V_{CT} value was obtained by chance, we recalculated V_{CT} 1000 times for each gene after random permutations of the ecotype labels. The 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 and Hochberg 1995). V_{CT} with corrected p-values smaller than 0.05 were considered significantly differentiated between lake and river ecotypes. For candidate gene eCNVs, we also calculated V_{CT} between each population pair, in the same way as V_{CT} was calculated for all populations combined but without population structure in the ANOVA model.

Allelic differentiation of eSNPs

In addition to the evaluation of copy number differentiation, we calculated nucleotide differentiation between lake and river ecotypes for each SNP identified as eSNPs, evaluated as F_{CT} using the locus-by-locus AMOVA approach implemented in Arlequin (Excoffier and Lischer 2010). The F_{CT} was calculated as the percentage of variance between groups (lake versus river ecotypes) relative to the total variance, using a hierarchical structure that groups lake and river ecotypes into 4 populations each. This AMOVA approach provides a more sensitive way to qualitatively evaluate habitat-specific patterns across replicated population pairs, compared to methods that scan for outlier regions in each population pair separately to identify parallel regions based on shared outliers (e.g. Feulner, et al. 2015), for the same reason as mentioned above for V_{CT}. We used permutation tests implemented in Arlequin to determine the significance of the F_{CT} values and identify eSNPs with significant F_{CT} values ($p<0.05$ from 1023 permutations).

Identifying correlations between expression and genetic differentiation

A genome-wide correlation between gene expression differentiation (P_{CT}) and genetic differentiation (V_{CT} and F_{CT}) was performed on all expressed genes. For this analysis, V_{CT} was calculated for each of 350 gene CNVs that had expression (not only gene eCNVs), and F_{CT} was calculated for each of 11,935 autosomal protein-coding genes that had expression (not only for eSNPs), excluding genes in CNV regions. F_{CT} was evaluated for each gene based on SNPs in the 5kb upstream regions, using the AMOVA approach implemented in Arlequin (Excoffier and Lischer 2010). With the resulting matrixes of P_{CT}, V_{CT} and F_{CT} of all genes expressed in the head kidney and/or spleen, the Spearman's rank correlation was used to test for correlation in each tissue between (a) P_{CT} and FV_{CT} and between (b) P_{CT} and F_{CT}. All statistical analyses were carried out using the package R version 3.0.2 (R Development Core Team 2011) unless otherwise indicated.

Testing for gene ontology enrichment in genes with eSNPs and eCNVs

We tested for enrichment of gene ontology (GO) terms among the gene eCNVs, the genes with eSNPs, the gene eCNVs with significant V_{CT}, and the genes with eSNPs with significant F_{CT}. The enrichment tests were conducted with topGO (Alexa and Rahnenfuhrer 2016), based on Fisher's exact tests applying Benjamini-Hochberg's multiple-test correction. We used different background gene sets depending on the enrichment analysis: we compared gene eCNVs to all expressed genes in either tissue and to all gene CNVs that are expressed in either tissue; we compared genes with eSNPs to all genes that were included in the eSNP tests; we compared gene eCNVs with significant V_{CT} to the set of gene eCNVs; we compared genes with eSNPs with significant F_{CT} to all genes with eSNPs. Overrepresented GO terms were those with corrected p-values smaller than 0.05.

Results

We first evaluated genotype-expression relationships using CNVs and SNPs, and then investigated whether they contribute to expression divergence between ecotypes. Our overarching goal was to evaluate the relationship between genetic differentiation of the two variant types and gene expression differentiation between replicated pairs of lake and river three-spined stickleback ecotypes.

Gene copy number and expression levels are largely positively correlated

Out of a total of 19,782 protein-coding autosomal genes in stickleback genome, we identified 832 gene CNVs among our samples. Among these gene CNVs, 350 CNVs had available gene expression data, out of which 140 (40%) 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), *slc47a1* (solute carrier family 47, member 1, ENSGACG00000020614) and two uncharacterized genes (ENSGACG00000020469 and ENSGACG00000012806) in head kidney samples, as well as *cyp3c1* (cytochrome P450 family 3 subfamily A member 43, ENSGACG00000010952) in spleen samples. The other 135 genes (39% of all expressed gene CNVs) had a positive correlation in at least one of the two tissues and were considered "gene eCNVs" (Figure 1, Table S2). Amongst these 135 gene eCNVs, 10 were only expressed and had a positive correlation in one tissue (either head kidney or spleen), 65 were expressed in both tissues while the expression was correlated with copy number in one tissue, while 60 were expressed in both tissues and showed a positive correlation between copy number and expression in both. Among

the genes that were expressed in either the head kidney or spleen tissues, gene eCNVs were enriched for antigen processing and presentation (GO:0019882, with 4 out of 28 genes), immune response (GO:0006955, with 5 out of 72 genes), major histocompatibility complex (MHC) protein complex (GO:0042611 with 4 of 27 genes) and MHC class I protein complex (GO:0042612 with 4 of 18 genes). MHC immune genes were amongst functional categories that were previously reported as enriched among all gene CNVs in sticklebacks (Chain, et al. 2014). When comparing gene eCNVs against all gene CNVs that were expressed in either tissue, there was no GO term enrichment observed.

Ten eCNVs show copy number differentiation between ecotypes

As gene eCNVs are the putative genetic variants that affect gene expression, we evaluated differentiation in their gene copy numbers between ecotypes, which could contribute to gene expression divergence. We estimated V_{CT} for each gene eCNV, which is the relative variance in gene copy numbers between ecotypes compared to the variance within ecotypes. Out of a total of 135 gene eCNVs, 10 (7.4%) had a significant V_{CT} (FDR<0.05, permutation test), with V_{CT} values ranging from 0.144 to 0.578 (Table 1). Of these ten genes, seven have higher average copy numbers in lake ecotypes than in river ecotypes, and three have higher copy numbers in river ecotypes. The 10 gene eCNVs with significant V_{CT} are distributed across six of 20 stickleback autosomes (Figure 2a). The GO annotations of the ten V_{CT} significant genes show that they are associated with various functions including ion binding, GTP binding, peptidase activity, diphosphatase activity and transmembrane transport (Table 1). But there was no functional enrichment of the ten gene eCNVs with significant V_{CT} compared to all gene eCNVs.

An abundance of genes with SNPs in *cis* are associated with expression

In addition to the CNVs associated with gene expression, we also investigated SNPs that are associated with gene expression. Out of a total of 870,917 SNPs within 5kb range of the TSSs of 11,360 genes expressed in either tissues, 8,353 SNPs were found associated with expression of 1,351 genes in the head kidney, 4,261 SNPs associated with expression of 746 genes in the spleen, including 1,336 SNPs associated with expression of 232 genes in both tissue types (corrected p-values < 0.05, Table S3). In total, 11,278 SNPs associated with 1,865 genes were determined as eSNPs that putatively contribute to gene expression differences among individuals. These eSNPs are symmetrically distributed across the 5 kb upstream and downstream range, with a slight peak within the 1kb range of the TSSs (Table S3). No GO term was enriched for the genes with eSNPs when compared to the joint set of 11,360 genes tested in the eSNPs analyses.

Fourteen eSNPs show allelic differentiation between ecotypes

For each eSNP, we evaluated the nucleotide differentiation, F_{CT}, between lake and river ecotypes. We found that 90.9% of eSNPs had negative or zero F_{CT} values, indicating no differentiation between lake and river fish populations. Out of the 1,112 eSNPs with a positive F_{CT}, 14 were significantly differentiated ($p<0.05$, permutation test), with F_{CT} values ranging from 0.120 to 0.378 (Figure 2a). These 14 eSNPs were associated with expression of 14 different genes. These 14 genes are annotated with various functions spanning mRNA splicing, DNA binding, rRNA methylation, signal transduction, ATP binding and GTP binding (Table 2), with no significant enrichment of GO categories compared to the set of genes with eSNPs.

One eSNP and two eCNVs display expression differentiation between ecotypes

The eSNPs and the gene eCNVs that are differentiated between ecotypes putatively contribute to expression differentiation. Amongst 12,105 genes expressed in the head kidney and 12,451 genes in the spleen, we identified 115 and 88 genes with significant P_{CT} respectively (FDR<0.05, Table S4). Out of these genes, we found one gene with significant P_{CT} (0.217) in the head kidney that also had an eSNP with significant F_{CT} (Figure 2). The P_{CT} in the spleen was 0.142 (FDR=0.11). The gene is dehydrogenase/reductase (SDR family) member 13a, duplicate 3 (*dhrs13a.3*, ENSGACG00000013614), a carboxypeptidase that catalyzes hydrolysis of peptide bonds (Uniprot entry: G3PTQ4). The SNP residing 630 bp upstream of the TSS of this gene had a F_{CT} value of 0.204, and was significantly associated with gene expression in both tissues. We also found two genes with significant P_{CT} that exhibited both differentiation in copy numbers (significant V_{CT}) and significant correlations between gene copy number and gene expression (gene eCNVs) in both tissues. The gene *cathepsin A* (ENSGACG00000015897) had significant P_{CT} in spleen (0.289; P_{CT} of 0.159 in head kidney) and the highest V_{CT} (0.578) amongst all gene CNVs (Figure 2). The other gene, GTPase, IMAP family member 7 (GIMAP7, ENSGACG00000018877), had significant P_{CT} identified in head kidney (0.245; P_{CT} of 0.184 in spleen) and a V_{CT} of 0.348 (Figure 2).

<u>eSNP regulating expression differentiation in *dhrs13a.3*</u>

Examining the differentiation signals within each population pair, the gene *dhrs13a.3* had higher expression levels in a subset of lake populations: in the head kidney of G1 (P_{CT} =0.648) and G2 (P_{CT} =0.204) but not in No (negative P_{CT}) and Ca (P_{CT} =0.076) (Figure 3c); in the spleen of G1 (P_{CT} =0.305) and No (P_{CT} =0.184) but not in G2 (P_{CT} =0.076) and Ca (negative P_{CT}). The genotypes of the eSNP residing 630 bp upstream of the TSS of *dhrs13a.3* were significantly correlated with gene expression levels across individuals in

both tissue types (FDR<0.001, Figure 3b showed in head kidney). This SNP was differentiated between lake and river ecotypes and had consistently higher allele frequency of the allele G in the lake populations (fixed in G1_L and G2_L, and 83.3% in No_L and Ca_L) and higher allele frequency of T in the river populations (25% in G1_R, and 41.7% in G2_R, No_R and Ca_R, Figure 3a). Both alleles occur in all four population pairs, and we confirmed that both were also present in a marine population from the North Sea (Feulner et al. 2012), with a low frequency of the T allele (8.3%). This suggests that the T allele derives from standing genetic variation in the ancestral marine populations, and repeatedly increased in frequency among river populations possibly due to positive selection. However, no selective sweep was found based on nucleotide diversity ($\boldsymbol{\pi}$) in the 50kb flanking region of the SNP, which did not differ between lake and river populations (Figure 4a). The gene region of *dhrs13a.3* harbors 51 SNPs across the four population pairs, with two synonymous and two non-synonymous SNPs in the exons, and other SNPs in the introns. The non-synonymous SNP, which substitutes a glycine with an arginine in the first exon, has the minor allele present in G1_L and G2_R with frequencies of 50% and 16.7%, respectively. The other non-synonymous SNP, which substitutes a cysteine with a phenylalanine in the third exon, has the minor allele present in Ca_L with a frequency of 16.7%.

<u>eCNV regulating expression differentiation in *cathepsin A*</u>

The gene *cathepsin A* had higher expression levels in spleen among river sticklebacks in the two German population pairs G1 (P_{CT} =0.664) and G2 (P_{CT} =0.409; Figure 3f), but was not differentially expressed in No nor Ca (negative P_{CT} values). In head kidney tissues, this gene also had higher expression in river sticklebacks in the population pairs of G1 (P_{CT} =0.797) and G2 (P_{CT} =0.190) and Ca (P_{CT} =0.112) whereas in No it had

higher expression in the lake fish (P_{CT} =0.521). The consistent differential expression in the two German population pairs was accompanied by copy number differentiation. This gene was the most differentiated gene CNV between lake and river sticklebacks in the two German population pairs (V_{CT} of 0.96 in G1 and 0.51 in G2) as previously reported (Chain, et al. 2014), but not differentiated in No nor Ca (V_{CT} =0) suggesting that the two German population pairs drive the overall habitat-specific signal (Figure 3e). We further identified *cathepsin A* as a gene eCNV, meaning that the gene copy numbers were significantly correlated with gene expression levels across individuals (FDR<0.001 in both tissue types, Figure 3d). To investigate whether the *cathepsin A* CNV is derived from standing genetic variation from an ancestral population, we searched for the presence of CNVs in an adjacent marine population from the North Sea (Feulner et al. 2012). The gene *cathepsin A* was not a CNV in the marine population, suggesting that the gene duplication occurred since the divergence of the freshwater populations (G1 and G2) from the marine population, or that the marine samples that were sequenced did not capture this variation. Note that the marine sampling only consists of 6 individuals, hence we lack power to detect variants at low frequency. A 5kb region in the gene region of *cathepsin A* was depleted with SNPs in G1_R leading to a nucleotide diversity ($\boldsymbol{\pi}$) of zero despite being duplicated compared to G1_L, suggesting a signature of background selection on the duplication (Figure 4b). In the other German populations, the gene harbors 23 SNPs, with two synonymous and one non-synonymous SNP. The non-synonymous SNP, which substitutes a leucine by a phenylalanine in an alternatively spliced exon, has the minor allele present as heterozygous in three individuals in G1_L and in two individuals in G2_L, and as homozygous in one G2_L individual.

<u>eCNV regulating expression differentiation in *GIMAP7*</u>

The gene *GIMAP7* had overall higher expression levels in the head kidney among lake ecotypes, and comparisons within population pairs found consistent directional differences across population pairs (Figure 3i). P_{CT} in the population pairs ranged from 0.11 in G1, to 0.19 in G2, and 0.39 in Ca whereas expression levels did not meet filtering criteria in No. The expression in spleen tissues displayed the same direction of expression changes between lake and river sticklebacks as in the head kidney, but differentiation was less pronounced: P_{CT} of 0.05 in G1, 0 in G2, 0.68 in No and 0.07 in Ca. The V_{CT} values were reasonably high in at least three population pairs: 0.53 in G1, 0.64 in No and 0.70 in Ca (Figure 3h). As with *cathepsin A, GIMAP7* was a gene eCNV (FDR=0.0074 in head kidneys and FDR<0.001 in spleen, Figure 3g). *GIMAP7* was not detected as a CNV in the North Sea marine population. This suggests independent duplication and deletion events in the freshwater populations since they diverged from the marine ancestor or that this variant is at low frequency in the marine population. In the genomic regions adjacent to GIMAP7, we found no differences in the levels of nucleotide diversity amongst the eight freshwater populations (Figure 4c). The gene region harbors a total of 38 SNPs across the four population pairs, 24 of which are non-synonymous. This suggests that duplication and deletion of this gene might also contribute to the amino acid sequence diversification across population pairs.

Genome-wide correlation between genetic differentiation and expression differentiation

Genome-wide, F_{CT} in *cis*-regulatory regions did not significantly positively correlate with P_{CT} in either head kidney or spleen (rho=0.011, p=0.12, n=10671 in head kidney and rho=0.006, p=0.24, n=10974 in spleen; one-sided Spearman rank correlation). V_{CT} had a significant positive correlation with P_{CT} in spleen but not in the head kidney

(rho=0.166, p<0.001 for spleen; rho=0.064, p=0.064 for head kidney; one-sided Spearman rank correlation).

Discussion

The genetic underpinnings of expression differentiation in adaptive evolution remain a focus of intense research. In this study, we combined genome-wide genetic variation and transcriptomic data from repeatedly evolved ecotypes of the three-spined stickleback to better understand their relationships in the process of adaptation to distinct habitats. We first report a prevalent dosage effect of CNV genes on gene expression and numerous SNPs in *cis* associated with expression. The prevalent association between genetic variants and expression levels might provide phenotypic variation that promotes adaptation to distinct lake and river habitats. We describe one gene with a differentiated SNP that is associated with expression differentiation between lake and river populations, and two genes with significant associations between copy number differentiation and expression differentiation. These findings provide evidence that both SNPs and CNVs contribute to gene expression differentiation between recently diverged ecotypes.

Dosage effects of CNVs contribute to expression differentiation

CNVs reflect components of genome architectures that vary in the number of copies of a sequence and have been proposed to have a greater impact on gene expression compared to sequence modifications (Sudmant, et al. 2015; Huddleston and Eichler 2016). We found that 39% (135) of all expressed gene CNVs have a positive association with expression in at least one of the two tissues sampled, with 60 gene CNVs showing significant positive association in both tissues. These results demonstrate prevalent

dosage effects on gene expression across tissue types. Similar number of genes show associations between CNVs and expression changes in humans (e.g. 110 genes in (Schlattl, et al. 2011) and 44-96 genes in (Stranger, et al. 2007)) and a similar proportion (42%) of genes in *Drosophila* (Cardoso-Moreira, et al. 2016). Recently, the Genotype-Tissue Expression (GTEx) Project also found large effect sizes of structural variations (SVs) on gene expression in humans and highlighted the likely causality of many CNVs (Chiang et al. 2017). This is consistent with our findings of 135 gene eCNVs as putative causal variants for expression variation. Whereas the 135 eCNVs are not enriched in any particular function compared to the whole set of gene CNVs, they are enriched for functions of antigen processing and genes of the adaptive immune system (MHC genes) compared to the genomic background. These two immune-related functional categories are a subset of enriched functions of gene CNVs overall (Chain, et al. 2014), suggesting that the immune system might be amenable to expression differentiation via copy number changes. It is plausible that immune-related gene CNVs play an important role in adaptation to different parasite pressure in their natural environments, and contribute to observed divergences between lake and river ecotypes (Eizaguirre, et al. 2009; Eizaguirre and Lenz 2010; Eizaguirre, et al. 2011).

The integration of differentiation patterns of gene copy numbers and gene expression amongst gene CNVs in the same individuals enabled us to investigate the dosage effects of CNVs in the context of ecotype divergence. However, there was a weak correlation between P_{CT} and V_{CT} genome-wide. This is consistent with work performed on *Drosophila* showing that the parallel differentiation of CNVs does not necessarily correlate with expression differentiation (Schrider, et al. 2016). These together indicate that not all CNVs affect expression, at least not in all tissues, and that mechanisms other than linear dosage effects are also relevant. For example, some gene CNVs can be dosage insensitive

(Zhou, et al. 2011), and others can affect gene expression through compensatory mechanisms (Henrichsen, et al. 2009). While not all CNVs are expected to contribute to population differentiation, the ones where copy numbers and expression are differentiated between habitat types are promising candidate genes involved in adaptation.

Genes underlying divergent adaptation should possess both high copy number differentiation (V_{CT}) and high expression differentiation (P_{CT}) between ecotypes, in addition to showing a positive correlation between copy numbers and gene expression levels (i.e. gene eCNVs). Here, we detected two genes, *cathepsin A* and *GIMAP7*, that fulfill both criteria and are therefore good candidates for being repeatedly driven by adaptive divergence between lake and river populations. The gene *cathepsin A* had the highest copy number differentiation amongst all gene CNVs and was present in more copies among the river ecotypes from the German populations than the German lake ecotypes, driving the overall differentiation signal. This gene encodes 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 roles in MHC class II antigen presentation (Hsing and Rudensky 2005). More copies of the gene and therefore higher expression conceivably impact the immune response, while most of the gene region is depleted from variation despite the duplication in G1_R, suggesting background selection on the duplication. As river sticklebacks have lower MHC diversity compared to lake ecotypes (Eizaguirre, et al. 2011), the higher copy number and expression of this gene potentially has a compensatory role and contributes to the defense against parasites specific to the river habitat. In contrast, lake ecotypes across population pairs were found to have higher copy numbers and higher expression of the gene *GIMAP7*, a GTPase that contains a domain AIG1-type G with immunity-associated functions (Krücken, et al. 2004;

Schwefel, et al. 2010). The increase in *GIMAP7* copy number is associated with higher expression, possibly contributing to higher immune competence in lake individuals, as the parasite pressure is more intense in lake habitats (Scharsack, et al. 2007; Eizaguirre, et al. 2011). The matching habitat-specific expression patterns of *cathepsin A* and *GIMAP7* in immune tissues add to previous findings that CNVs are likely an important source of genetic variation that can help shape the host innate and adaptive immune response (Chain, et al. 2014; Machado and Ottolini 2015). Our study on habitat-specific expression in immune tissues, which can potentially capture parasite-mediated selection, has revealed two immune-related gene CNVs associated with expression differentiation, whereas other CNVs possibly contribute to habitat-specific adaptations in other tissues not sampled in our study. Previous investigation between marine and freshwater sticklebacks identified 24 gene CNVs consistent with parallel evolution, two of which were also found with differential expression between photoperiod treatments (APOL2 and ENSGACG00000003408, Hirase, et al. 2014). These two genes were also gene CNVs in our population system, with ENSGACG00000003408 also marginally differentiated between our lake and river populations (V_{CT} = 0.124, FDR=0.053), but neither gene was expressed in our transcriptome data. In addition to Hirase, et al. (2014), our findings of two gene eCNVs with significant V_{CT} and P_{CT} highlight an important role of gene CNVs in adaptation to new environments in sticklebacks.

eSNPs in *cis* also contribute to expression variation

In addition to CNVs affecting gene expression, a total of 1,865 genes had SNPs in *cis*-regulatory regions identified as eSNPs putatively affecting gene expression. Though association tests between gene expression and SNPs do not necessarily reflect causal relationships, this result is consistent with previous studies that found abundant *cis*-

eQTLs associated with expression divergence between stickleback ecotypes (Ishikawa, et al. 2017; Pritchard, et al. 2017; Kitano, et al. 2018). Comparing marine and freshwater sticklebacks, Ishikawa, et al. (2017) reported that about half of their local eQTLs resided in genomic regions of high divergence. Extending the comparison to multiple population pairs and between lake and river populations, we identified a gene differentiated between ecotypes both at the genetic level of an eSNP and in gene expression. The lake and river sticklebacks used in this study exhibit low parallel genomic divergence despite an isolation-by-adaptation signal (Feulner, et al. 2015); genomic regions that most likely contribute to ecological divergence vary across different population pairs, suggesting the regulatory changes responsible for expression differentiation might also be population specific. As for *dhrs13a.3*, the homozygous T genotype of the eSNP 630 bp upstream of the TSS was associated with lower expression, and present in higher frequency in river populations where parasite abundance is generally much lower than in lakes (Scharsack, et al. 2007; Eizaguirre, et al. 2011). This allele is present in a detectable but low frequency (8.3%) in a source marine population (North See, Feulner et al. 2012) as well as in our lake populations, suggesting repeated increases in frequency in river habitats putatively due to habitat-specific adaptation.

Despite the abundance of genes with eSNPs, sequence differentiation of 5kb upstream regions had an overall non-significant correlation with expression differentiation. This lack of genome-wide correlation between sequence-based differentiation in *cis*-regulatory regions and expression differentiation is consistent with other studies in whitefish, flycatcher and *Drosophila* (Renaut, et al. 2012; Zhao, et al. 2015; Uebbing, et al. 2016), and can be at least partly explained by the narrow transcriptomic snapshot analyzed. Sequence differentiation might still impact expression differentiation in other tissues or at different developmental times not captured in our data. We also

cannot exclude the impact that environmental plasticity might play in shaping expression differentiation. While trans-regulatory changes may also contribute to expression divergence (e.g. Hart et al. 2018), we focused on *cis*-regulatory changes, which were found to account for large parts of parallel expression changes between marine and freshwater sticklebacks (Verta and Jones 2019). Taken together, our results highlight examples of SNPs and CNVs that contribute to expression differentiation linked to adaptive divergence.

Conclusion

By combining genome and transcriptome data from the same individuals across independently evolved population pairs, we describe generalities of the genetic basis of gene expression differentiation between lake and river sticklebacks. We revealed numerous changes of nucleotides in *cis*-regulatory elements that are associated with expression variation and prevalent dosage effects of CNVs on gene expression, providing variation that can foster rapid adaptation to different environments. We report one SNP in *cis* and two CNVs linked to gene expression differentiation that likely contribute to divergence between repeatedly evolved ecotypes. Our findings highlight both SNPs and CNVs as sources of genetic variation that promote repeated adaptation via cis-regulatory effect or dosage effect on gene expression.

Figures

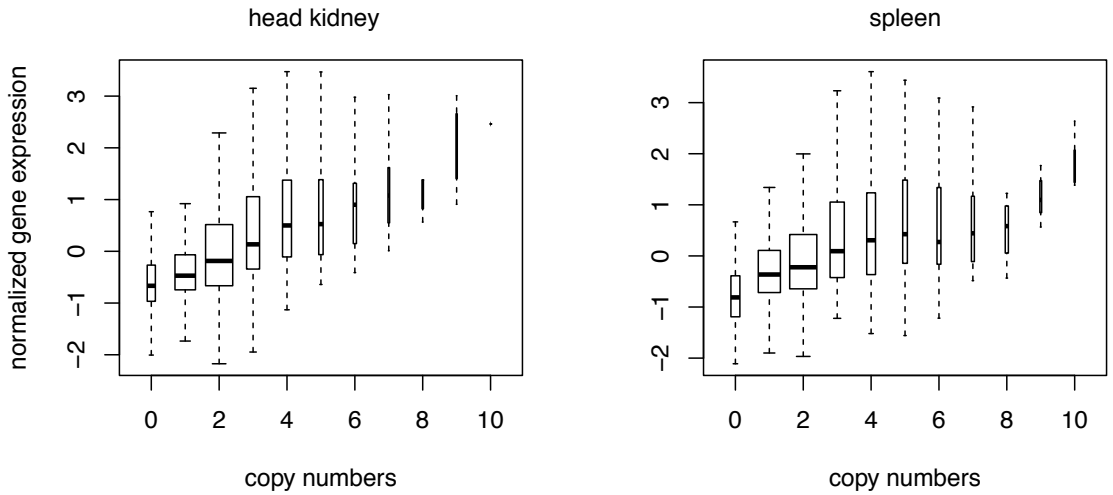

Figure 1. Normalized gene expression levels for a given gene copy number summarized across all gene eCNVs and individuals (n genes = 135; n individuals = 40 for head kidney and n = 36 for spleen). Expression levels were evaluated in head kidney and spleen separately. Expression levels of each gene were centered to zero and scaled by the standard deviations. The widths of boxes represent the relative sample size (i.e. number of genes in each copy number category). Only copy numbers up to 10 are shown.

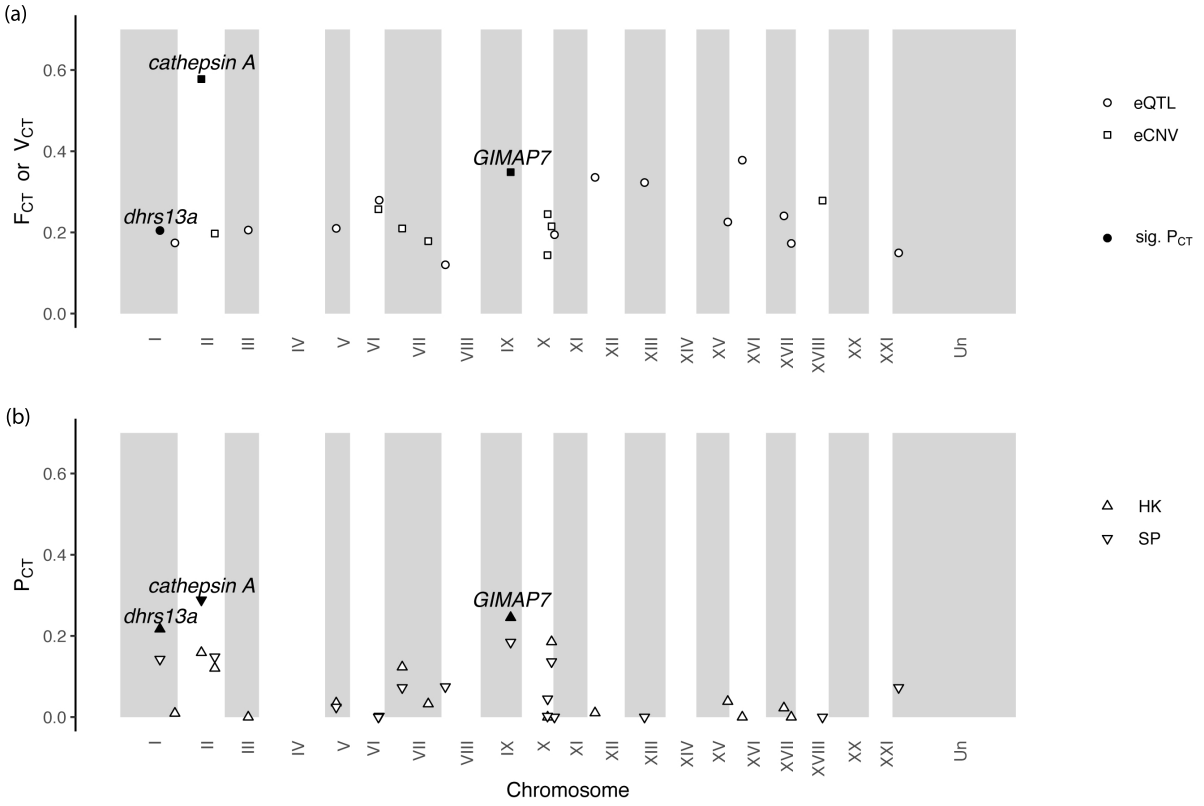

Figure 2. Genes with eSNPs with significant F_{CT} and eCNVs with significant V_{CT} between lake and river stickleback populations and P_{CT} of these same genes. (a) Genes with eSNPs with significant F_{CT} (circle) and gene eCNVs with significant V_{CT} (square) along the genome; and (b) P_{CT} of these same genes in the head kidney (triangle) and/or in the spleen (inverted triangle). Only the P_{CT} in the tissues where the eSNPs or the gene eCNVs were identified are shown. Genomic locations include twenty linkage groups of the stickleback genome representing autosomes (excluding the sex chromosome XIX), in addition to unplaced scaffolds (Un). The filled shapes indicate the three genes with significant P_{CT}.

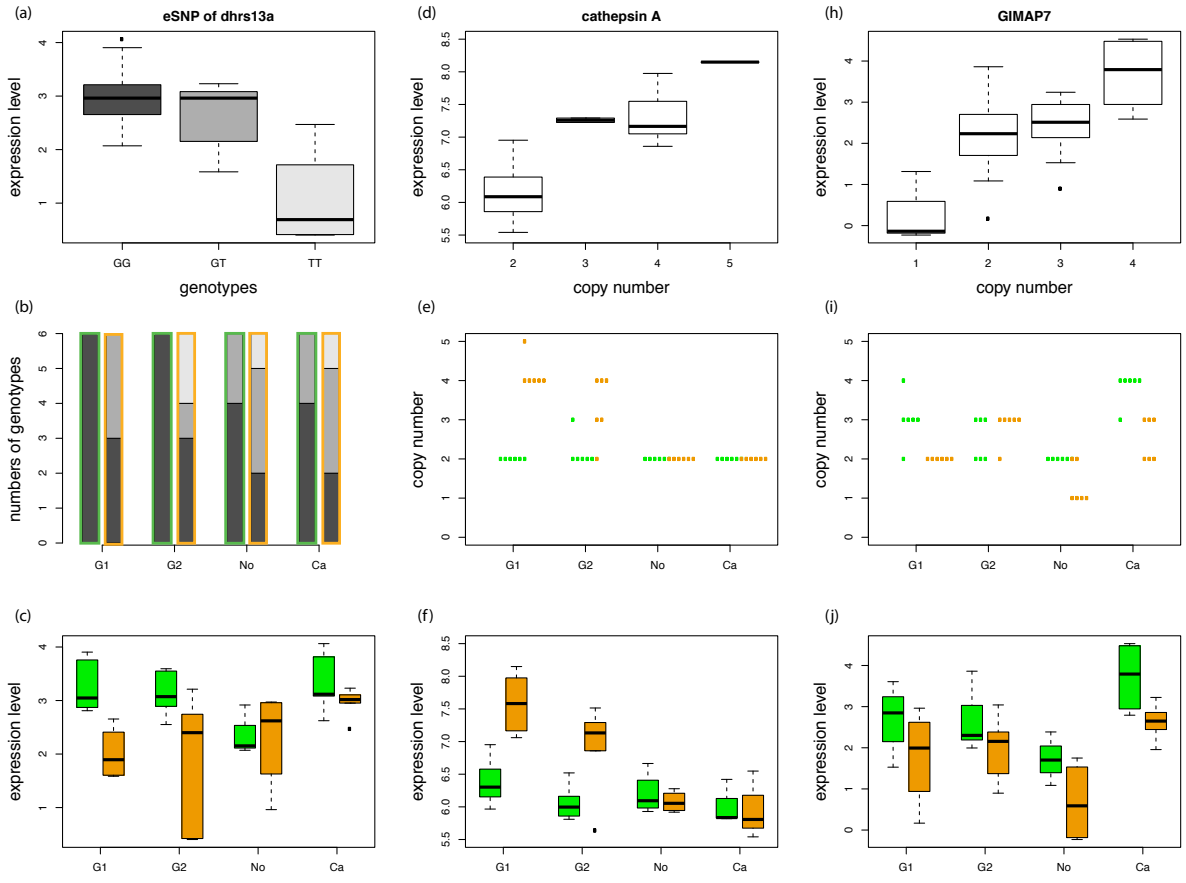

Figure 3. Gene *dhrs13a* with an eSNP with significant F_{CT} and gene eCNVs, *cathepsin A* and *GIMAP7*, with significant V_{CT} that also had significant P_{CT} between lake and river sticklebacks. (a) Association between eSNP genotypes and expression levels in the head kidney of *dhrs13a*, with y-axes indicating expression levels of the different genotypes in boxplots summarizing normalized read counts across individuals. (b) Genotypes of the eSNP across four population pairs (G1: Germany 1, G2: Germany 2, No: Norway, Ca: Canada) where the bars with green border represent lake populations and the bars with orange border represent river populations. The colours for the genotypes are the same as in (a). (c) Expression differences in the head kidney across the same individuals where lake populations indicated in green and river populations in orange. (d & g) The association between gene copy numbers and gene expression in *cathepsin A* in the spleen (d) and *GIMAP7* in the head kidney (g). (e & h) Habitat-specific patterns of gene copy

number of *cathepsin A* (e) and *GIMAP7* (h) across populations (dots represent lake and river individuals in green and orange, respectively). (f & i) the habitat-specific expression patterns of the same two genes, *cathepsin A* in the spleen (f) and *GIMAP7* in the head kidney (i) across populations.

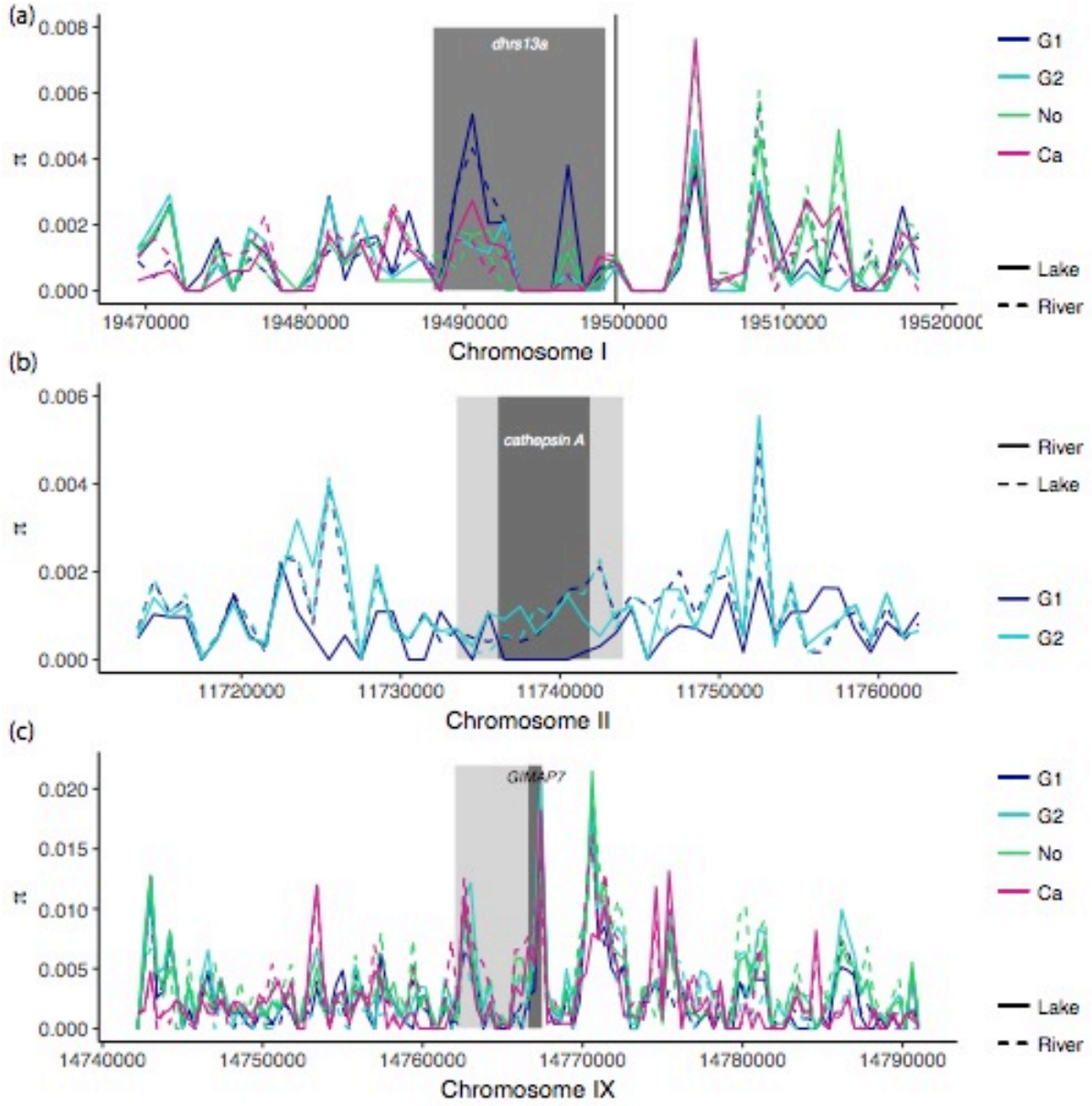

Figure 4. Nucleotide diversity ($\boldsymbol{\pi}$) in the 50kb flanking regions of the three candidate genes, the gene *dhrs13a* with an eSNP (a) and two eCNV genes, *cathepsin A* (b) and *GIMAP7* (c). In (a) the gene region is in dark grey and the eSNP denoted by a black vertical line. In (b) and (c) the gene regions are in dark grey and the CNV regions are in light grey. For the three genes, $\boldsymbol{\pi}$ was calculated for each population separately. For *dhrs13a* and *cathepsin A*, $\boldsymbol{\pi}$ was calculated for each 1kb window and for *GIMAP7* $\boldsymbol{\pi}$ was calculated for 400b window to adjust for SNP densities in each window. Solid lines represent populations with higher gene copy number (lake for *dhrs13a* and *GIMAP7* and river for *cathepsin A*)

whereas dashed lines represent populations with lower gene copy number. For *cathepsin A*, we focused on G1 and G2 population pairs because CNVs were identified in only these two population pairs.

Tables

Table 1. Genes with significant differentiation in gene copy numbers (V_{CT}) between lake and river ecotypes

Table 2. Genes with significant differentiation in eSNPs (F_{CT}) between lake and river ecotypes

Supplementary Material

Figure S1. The geographical map of the sampling sites. Adapted from Feulner et al. 2015.

Table S1. Accession numbers of transcriptome and genome samples.

Table S2. List of gene eCNVs showing V_{CT} and parameter estimates from linear mixed effect models between gene copy number and gene expression (in two tissue types separately).

Table S3. List of genes and the eSNPs showing associations between SNPs and gene expression (in two tissue types separately) and F_{CT} values.

Table S4. List of genes with significant P_{CT} in either of the two tissues.

Supplementary Information. R scripts for eCNV and V_{CT} analyses.

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. We thank Belinda Chang and three anonymous referees for their help in improving this manuscript.

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 on research design 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.

References

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 Res 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.

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

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.

Barrett RD, Hoekstra HE. 2011. Molecular spandrels: tests of adaptation at the genetic level. Nat Rev Genet 12:767-780.

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.

Berner D, Adams DC, Grandchamp AC, Hendry AP. 2008. Natural selection drives patterns of lake–stream divergence in stickleback foraging morphology. J Evol Biol 21:1653-1665.

Birchler JA, Veitia RA. 2012. Gene balance hypothesis: Connecting issues of dosage sensitivity across biological disciplines. Proceedings of the National Academy of Sciences of the United States of America 109:14746-14753.

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.

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 Res 26:787-798.

Carroll SB. 2008. Evo-Devo and an Expanding Evolutionary Synthesis: A Genetic Theory of Morphological Evolution. Cell 134:25-36.

Chain FJ, Feulner PG, Panchal M, Eizaguirre C, Samonte IE, Kalbe M, Lenz TL, Stoll M, Bornberg-Bauer E, Milinski M, et al. 2014. Extensive copy-number variation of young genes across stickleback populations. PLoS Genet 10:e1004830.

Chan YF, Marks ME, Jones FC, Villarreal G, Jr., Shapiro MD, Brady SD, Southwick AM, Absher DM, Grimwood J, Schmutz J, et al. 2010. Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a Pitx1 enhancer. Science 327:302-305.

Chiang C, Scott AJ, Davis JR, Tsang EK, Li X, Kim Y, Hadzic T, Damani FN, Ganel L, GTEx Consortium, Montgomery SB, Battle A, Conrad DF, Hall IM. 2017. The impact of structural variation on human gene expression. Nat Genet 49:692-699.

Ching T, Huang S, Garmire LX. 2014. Power analysis and sample size estimation for RNA-Seq differential expression. RNA 20:1684-1696.

Colombo M, Diepeveen ET, Muschick M, Santos ME, Indermaur A, Boileau N, Barluenga M, Salzburger W. 2013. The ecological and genetic basis of convergent thick-lipped phenotypes in cichlid fishes. Mol Ecol 22:670-684.

Deagle BE, Jones FC, Chan YF, Absher DM, Kingsley DM, Reimchen TE. 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.

Eizaguirre, C, & Lenz, TL. 2010. Major histocompatibility complex polymorphism: dynamics and consequences of parasite-mediated local adaptation in fishes. J Fish Biol, 77(9): 2023-2047.

Eizaguirre C, Lenz TL, Sommerfeld RD, Harrod C, Kalbe M, Milinski M. 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. Ecol Lett 12:5-12.

Excoffier L, Lischer HE. 2010. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. Mol Ecol Resour 10:564-567.

Faucon F, Dusfour I, Gaude T, Navratil V, Boyer F, Chandre F, Sirisopa P, Thanispong K, Juntarajumnong W, Poupardin R, et al. 2015. Identifying genomic changes associated with insecticide resistance in the dengue mosquito Aedes aegypti by deep targeted sequencing. Genome Res 25:1347-1359.

Feulner PG, Chain FJ, Panchal M, Huang Y, Eizaguirre C, Kalbe M, Lenz TL, Samonte IE, Stoll M, Bornberg-Bauer E, et al. 2015. Genomics of divergence along a continuum of parapatric population differentiation. PLoS Genet 11:e1004966.

Gamazon ER, Stranger BE. 2015. The impact of human copy number variation on gene expression. Brief Funct Genomics 14:352-357.

Gibson G. 2008. The environmental contribution to gene expression profiles. Nat Rev Genet 9:575-581.

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.

Hanson D, Hu J, Hendry AP, Barrett RDH. 2017. Heritable gene expression differences between lake and stream stickleback include both parallel and antiparallel components. Heredity 119:339-348.

Haraksingh RR, Snyder MP. 2013. Impacts of variation in the human genome on gene regulation. J Mol Biol 425:3970-3977.

Hart J, Ellis NA, Eisen MB, Miller CT. 2018. Convergent evolution of gene expression in two high-toothed stickleback populations. PLoS Genet 14(6):1007443.

Hartasánchez DA, Brasó-Vives M, Heredia-Genestar JM, Pybus M, Navarro A. 2018. Effect of collapsed duplications on diversity estimates: what to expect. Genome Biol Evol 10(11): 2899-2905

Henrichsen CN, Chaignat E, Reymond A. 2009. Copy number variants, diseases and gene expression. Hum Mol Genet 18:R1-8.

Hirase S, Ozaki H, Iwasaki W. (Hirase2014 co-authors). 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.

Hsing LC, Rudensky AY. 2005. The lysosomal cysteine proteases in MHC class II antigen presentation. Immunol Rev 207:229-241.

Huang Y, Chain FJJ, Panchal M, Eizaguirre C, Kalbe M, Lenz TL, Samonte IE, Stoll M, Bornberg-Bauer E, Reusch TBH, et al. 2016. Transcriptome profiling of immune tissues reveals habitat-specific gene expression between lake and river sticklebacks. Mol Ecol 25:943-958.

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

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.

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. Convergence and non-convergence in ecological, phenotypic, and genetic divergence across replicate population pairs of lake and stream stickleback. Evolution; international journal of organic 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.

Katju V, Bergthorsson U. 2013. Copy-number changes in evolution: rates, fitness effects and adaptive significance. Frontiers in Genetics 4:273.

Kitano J, Ishikawa A, Kusakabe M. 2019. Parallel transcriptome evolution in stream threespine sticklebacks. Develop. Growth Differ. 61:104–113.

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.

Leder EH, McCairns RJ, Leinonen T, Cano JM, Viitaniemi HM, Nikinmaa M, Primmer CR, Merila J. 2015. The evolution and adaptive potential of transcriptional variation in sticklebacks--signatures of selection and widespread heritability. Mol Biol Evol 32:674-689.

Lenz TL. 2015. Transcription in space – environmental vs. genetic effects on differential immune gene expression. Mol Ecol 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. Mol Ecol 22:774-786.

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.

Lucek K, Sivasundar A, Kristjánsson BK, Skúlason S, Seehausen O. 2014. Quick divergence but slow convergence during ecotype formation in lake and stream stickleback pairs of variable age. J Evol Biol 27:1878-1892.

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

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

McKinnon JS, Rundle HD. 2002. Speciation in nature: the threespine stickleback model systems. Trends in Ecology & Evolution 17:480-488.

Nguyen D-Q, Webber C, Ponting CP. 2006. Bias of Selection on Human Copy-Number Variants. PLOS Genetics 2:e20.

Ongen H, Buil A, Brown AA, Dermitzakis ET, Delaneau O. 2016. Fast and efficient QTL mapper for thousands of molecular phenotypes, *Bioinformatics*, 32(10) 1479–1485

Papp B, Pal C, Hurst LD. 2003. Dosage sensitivity and the evolution of gene families in yeast. Nature 424:194-197.

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.

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.

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.

R Development Core Team. 2011. R: A Language and Environment for Statistical Computing.

Ravinet M, Prodöhl PA, Harrod C. 2013. Parallel and nonparallel ecological, morphological and genetic divergence in lake–stream stickleback from a single catchment. J Evol Biol 26:186-204.

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. Philos Trans R Soc Lond B Biol Sci 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. Mol Ecol 10:2435-2445.

Rice AM, McLysaght A. 2017. Dosage sensitivity is a major determinant of human copy number variant pathogenicity. Nat Commun 8:14366.

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 Biol 11:R25.

Roesti M, Hendry AP, Salzburger W, Berner D. 2012. Genome divergence during evolutionary diversification as revealed in replicate lake-stream stickleback population pairs. Mol Ecol 21:2852-2862.

Scharsack JP, Kalbe M, Harrod C, Rauch G. 2007. Habitat-specific adaptation of immune responses of stickleback (Gasterosteus aculeatus) lake and river ecotypes. Proc Biol Sci 274:1523-1532.

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 Res 21:2004-2013.

Schrider DR, Hahn MW, Begun DJ. 2016. Parallel Evolution of Copy-Number Variation across Continents in Drosophila melanogaster. Mol Biol Evol 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.

Stamatoyannopoulos JA. 2004. The genomics of gene expression. Genomics 84:449-457.

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

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. Nat Ecol Evol 1:158.

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.

Tarazona S, García-Alcalde F, Dopazo J, Ferrer A, Conesa A. 2011. Differential expression in RNA-seq: A matter of depth. Genome Research 21:2213-2223.

Timur ZK, Akyildiz Demir S, Seyrantepe V. 2016. Lysosomal Cathepsin A Plays a Significant Role in the Processing of Endogenous Bioactive Peptides. Front Mol Biosci 3:68.

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. Mol Ecol 25:2015-2028.

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

Verta JP, Jones FC. 2019. Predominance of cis-regulatory changes in parallel expression divergence of sticklebacks. eLife 8:e43785

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

Whitehead A, Crawford DL. 2006b. Variation within and among species in gene expression: raw material for evolution. Mol Ecol 15:1197-1211.

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

Wittkopp PJ, Kalay G. 2011. Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. Nat Rev Genet 13:59-69.

Zhang F, Gu W, Hurles ME, Lupski JR. 2009. Copy Number Variation in Human Health, Disease, and Evolution. Annu Rev Genomics Hum Genet 10:451-481.

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 Genet 11:e1005184.

Zhou J, Lemos B, Dopman EB, Hartl DL. 2011. Copy-Number Variation: The Balance between Gene Dosage and Expression in Drosophila melanogaster. Genome Biol Evol 3:1014-1024.

Table 1. Genes with significant differentiation in gene copy numbers (V_{CT}) between lake and river ecotypes

Gene ID	Gene name	GO function			Tissue of eCNV	Higher copy number	V_{CT}	P_{CT}	
		Cellular component	Molecular function	Biological process				HK	SP
ENSGACG00000008264	novel gene	unknown			both	river	0.245	-0.025	0.045
ENSGACG00000010952	cytochrome P450 family 3 subfamily A member 43 (CYP3A43)	membrane; integral component of membrane	monooxygenase activity; iron ion binding; oxidoreductase activity; oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen; heme binding; metal ion binding	oxidation-reduction process	SP	lake	0.257	-0.024	-0.019
ENSGACG00000012073	novel gene	unknown			SP	lake	0.278	NA	-0.021
ENSGACG00000015897	cathepsin A	unknown	peptidase activity; serine-type carboxypeptidase activity; hydrolase activity	proteolysis	both	river	0.578	0.159	0.289 *

ENSGACG00000016770	deoxyuridine triphosphatase (dut)	unknown	dUTP diphosphatase activity	dUTP metabolic process	both	lake	0.197	0.120	0.148
ENSGACG00000018877	GTPase, IMAP family member 7 (GIMAP7)	unknown	GTP binding	unknown	both	lake	0.348	0.245 *	0.184
ENSGACG00000019933	si:dkey-85k7.12	unknown	GTP binding	unknown	both	lake	0.210	0.123	0.072
ENSGACG00000020614	solute carrier family 47 (slc47a1)	membrane; integral component of membrane	drug transmembrane transporter activity; antiporter activity	transmembrane transport	HK	lake	0.178	0.032	-0.001
ENSGACG00000008242	novel gene	unknown			SP	river	0.144	0.082	0.002
ENSGACG00000009551	ring finger protein 139 (rnf139)	membrane; integral component of membrane	zinc ion binding; metal ion binding,	unknown	both	lake	0.215	0.186	0.136

*: significant P_{CT}

NAs in P_{CT}: expression levels did not meet the filtering requirements and therefore P_{CT} were not calculated.

Table 2. Genes with significant differentiation in eSNPs (F_{CT}) between lake and river ecotypes

Gene ID	Gene name	GO function			eSNP			F_{CT}	P_{CT}	
		Cellular component	Molecular function	Biological process	position in relation to TSS	distance to TSS (bp)	Tissue		HK	SP
ENSGACG00000000642	novel gene	unknown			up-stream	1637	SP	0.150	NA	0.073
ENSGACG00000002647	alkB homolog 6 (alkbh6)	unknown	oxidoreductase activity	oxidation-reduction process	down-stream	3811	HK	0.378	-0.021	-0.005
ENSGACG00000003827	si:ch73-14h10.2	P-body; nucleus; spliceosomal complex; U6 snRNP; U4/U6 x U5 tri-snRNP complex	mRNA splicing, via spliceosome; mRNA processing; RNA splicing	RNA binding	up-stream	3063	SP	0.120	NA	0.075
ENSGACG00000004256	activity-dependent neuroprotector homeobox b (adnpb)	nucleus	nucleic acid binding; DNA binding	erythrocyte maturation	down-stream	2315	HK	0.336	0.010	-0.028
ENSGACG00000004442	myocardin related transcription factor Ba (mrtfba)	unknown			down-stream	1161	both	0.210	0.035	0.024

ENSGACG00000004844	RAS related (rras)	membrane	nucleotide binding; GTPase activity; GTP binding	signal transduction; Notch signaling pathway; maintenance of epithelial cell apical/basal polarity	down-stream	1404	SP	0.194	0.001	-0.006
ENSGACG00000009164	Ly1 antibody reactive homolog (lyar)	unknown	DNA binding	unknown	down-stream	1080	HK	0.241	0.023	0.012
ENSGACG00000009941	DIM1 dimethyladenosine transferase 1-like (dimt1l)	unknown	rRNA (adenine-N6,N6-)-dimethyltransferase activity; RNA binding; methyltransferase activity; rRNA methyltransferase activity; transferase activity	rRNA modification; rRNA processing; rRNA methylation; methylation	down-stream	3438	SP	0.323	-0.004	-0.028
ENSGACG00000011156	sulfotransferase family, cytosolic, 6b, member 1 (sult6b1)	unknown	sulfotransferase activity; transferase activity	cellular response to xenobiotic stimulus	down-stream	2520	SP	0.279	0.230	0.002
ENSGACG00000011426	si:ch73-267c23.10	membrane; integral	unknown	unknown	down-stream	1570	HK	0.173	-0.012	0.038

		component of membrane								
ENSGACG00000013118	B cell CLL/lymphoma 11B (BCL11B)	none	nucleic acid binding	unknown	up-stream	2912	HK	0.226	0.039	0.047
ENSGACG00000013614	dehydrogenase/reductase (SDR family) member 13a, duplicate 3 (dhrs13a.3)	unknown			up-stream	630	both	0.205	0.217 *	0.142
ENSGACG00000015279	downstream neighbor of SON (DONSON)	unknown			up-stream	1525	HK	0.174	0.009	0.268
ENSGACG00000016707	ATP-binding cassette, sub-family A (ABC1), member 4b (abca4b)	membrane; integral component of membrane	nucleotide binding; ATP binding; ATPase activity; ATPase activity, coupled to transmembrane movement of substances	transmembrane transport	down-stream	2738	HK	0.206	-0.008	-0.002

*: significant P_{CT}

NAs in P_{CT}: expression levels did not meet the filtering requirements and therefore P_{CT} were not calculated.