



Schneider, Kevin (2020) *Population genomics and molecular evolution in salmonids*. PhD thesis.

<https://theses.gla.ac.uk/81931/>

Copyright and moral rights for this work are retained by the author

A copy can be downloaded for personal non-commercial research or study, without prior permission or charge

This work cannot be reproduced or quoted extensively from without first obtaining permission from the author

The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author

When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given

Enlighten: Theses

<https://theses.gla.ac.uk/>
research-enlighten@glasgow.ac.uk

Population Genomics and Molecular Evolution in Salmonids

Kevin Schneider, BSc MSc



Submitted in fulfilment of the requirements for the
Degree of Doctor of Philosophy
Institute of Biodiversity, Animal Health and
Comparative Medicine
College of Medical, Veterinary and Life Sciences
November 2020

Abstract

Salmonid fishes are characterised by a very high level of variation in trophic, ecological, physiological, and life history adaptations to diverse environments. Some salmonid taxa show exceptional potential for fast, within-lake diversification into morphologically and ecologically distinct variants, often in parallel; these are the lake-resident charr and whitefish (several species in the genera *Salvelinus* and *Coregonus*). Such rapid diversifications can offer clues about the genetic drivers of ecological speciation by natural selection. The often high parallelism in the ecological adaptations and diversifications in *Salvelinus* and *Coregonus* species also makes it possible to assess what genetic factors are shared in lake divergences as well as amongst species of these two highly diversifying salmonids. Such sharing is often assessed by examining the overlap of signatures of natural selection across populations or species. Moreover, recent advances in ecological genomics have shown a role of introgression in distributing these shared adaptive genetic variants amongst populations.

First, to infer similarities of selection signals and shared selected genes and gene functions that potentially promote the diversification potential of species, I compared the signatures of selection in the sequences of transcriptomes of the diversifying salmonids *Salvelinus* and *Coregonus* against non-diversifying salmonids. Second, I analysed the influence the often complex demographic history of diverging species has on signals of positive selection at the population level using computer simulations. Third, informed by step two, I examined the sharing of selection across five divergences of *Salvelinus* into benthic and pelagic ecomorphs using whole-genome resequencing. Fourth, I analysed the role of adaptive introgression in promoting diversification and local adaptation in one exemplary population of *Salvelinus* based on whole-genome sequences.

I found evidence for limited but putatively ecologically significant sharing of genes under selection and higher sharing of gene functions and large-scale genomic regions under selection, both at the interspecies level in transcriptomes and at the interpopulation, whole-genome level in *Salvelinus*. The shared targets of selection in *Salvelinus* matched known quantitative trait loci of ecological importance in salmonids, with a predominance of genetic variants in putative regulatory regions. In an exemplary population of *Salvelinus*, putative adaptive introgression was shown to play a potential role in rapid diversifications. The results of this thesis support the notion that parallelism at the genetic level does not have to be pervasive in order for pronounced phenotypic parallelism to occur, implying a highly polygenic basis of most ecological adaptations.

Table of Contents

Abstract	2
Preface	9
Acknowledgements	10
Author's declaration	11
Chapter 1: General introduction	12
1.1 The origins of biological diversity	13
1.2 Evolutionary research based on genome-scale data	15
1.3 Methods for analysing genome-scale data and population genetic signatures	16
1.4 Rapid diversifications as models for the study of diversity evolution	17
1.5 Salmonids and Arctic charr as models for ecological and evolutionary genomics	17
1.6 Thesis outline	19
Chapter 2: Parallel selection on ecologically relevant gene functions in the transcriptomes of highly diversifying salmonids	21
2.1 Abstract	21
2.2 Introduction	22
2.3 Materials & methods	25
2.3.1 Taxonomic dataset	25
2.3.2 Orthogroup and OG tree inference	27
2.3.3 Selection analyses	28
2.3.4 Gene ontology analyses	29
2.3.5 Gene set enrichment analysis	30
2.4 Results	31
2.4.1 Selection parameter distribution and number of orthogroups under selection	31
2.4.2 Gene functions under relaxed selection	33
2.4.3 Gene functions under intensified selection	35
2.4.4 Gene functions under diversifying selection	37
2.4.5 Overlap between selection types	37
2.5 Discussion	39
2.5.1 Selection parameter distribution – shift to relaxation?	40
2.5.2 Gene functions under relaxed selection in <i>Coregonus</i> and <i>Salvelinus</i>	41
2.5.3 Gene functions under intensified selection in <i>Coregonus</i> and <i>Salvelinus</i>	41
2.5.4 Gene functions under diversifying selection in <i>Coregonus</i> and <i>Salvelinus</i>	42
2.5.5 Overlap between selection types	42
2.5.6 Whole-genome duplication and potential effects on relaxed selection	44
2.6 Conclusions	45
Chapter 3: Pitfalls and virtues of population genetic summary statistics: detecting selective sweeps in recent divergences	46
3.1 Abstract	46
3.2 Introduction	46
3.3 Materials & methods	50
3.3.1 Simulations	50
3.3.2 Summary statistics	54
3.3.3 BGD, a beta diversity measure using genetic distances derived from information theory	56
3.3.4 Accuracies of sweep prediction	57
3.4 Results	57
3.4.1 Patterns of selective sweeps across the genome	57
3.4.2 Accuracy in identifying sweep location across scenarios	58
3.4.3 Prediction of sweep location – scenario-dependent performance	60

	4
3.5 Discussion	62
3.5.1 The pros and cons of different metrics	62
3.5.2 The influence of demography, recombination, and sweep type	63
3.5.3 The utility of differentiation-based metrics	65
3.5.4 Potential caveats	66
3.5.5 Suggested steps for locating selective sweeps	67
Chapter 4: Shared genomic targets of selection in repeated sympatric divergences of Arctic charr	70
4.1 Abstract	70
4.2 Introduction	70
4.3 Materials & methods	72
4.3.1 DNA extraction & testing of library preparation methods	72
4.3.2 Library preparation & whole-genome sequencing	73
4.3.3 Read filtering, mapping, & calculation of genotype likelihoods	73
4.3.4 Population structure & admixture analyses	75
4.3.5 ABBA-BABA tests for introgression	75
4.3.6 Metrics of diversity & selection across the genome	76
4.3.7 Shared outlier genes & windows	77
4.3.8 Functional enrichment	77
4.3.9 Overlap of outlier genes with salmonid QTLs	78
4.3.10 Types of substitutions in outlier genes	78
4.3.11 Phylogenetic relationships of <i>grid</i> genes in Arctic charr	79
4.3.12 Principal component analysis on metrics & their correlations	79
4.4 Results	80
4.4.1 Mapping success and variant statistics	80
4.4.2 Population structure & admixture	80
4.4.3 Shared selection & differentiation outliers in benthic-pelagic ecomorph pairs	82
4.4.4 Phylogenetic relationships of <i>grid</i> genes	84
4.4.5 Nucleotide substitutions around outlier genes	84
4.4.6 Relationships between population genetic metrics	84
4.4.7 Functional overrepresentation in genomic regions of shared selection	85
4.4.8 Overlap of outlier genes with ecologically relevant QTLs	86
4.5 Discussion	88
4.5.1 Relationships among divergences & origin of selected genetic variation	88
4.5.2 Sharing of selection outliers reveals shared ecological challenges	90
4.5.3 Hot spots of differentiation & sharing of selected gene functions	92
4.6 Conclusions	94
Chapter 5: Adaptive introgression after secondary contact facilitates ecomorph diversity in Arctic charr	95
5.1 Abstract	95
5.2 Introduction	95
5.3 Materials & methods	98
5.3.1 Library preparation & whole-genome sequencing	98
5.3.2 Genotype calling, SNP calling, & SNP filtering	99
5.3.3 Population structure & four-taxon tests for introgression	100
5.3.4 Selection scans	101
5.3.5 Inference of adaptive introgression	101
5.3.6 Functional enrichment	102
5.4 Results	103
5.4.1 Mapping success and number of SNPs	103
5.4.2 Population structure	103
5.4.3 Gene functions under selection	104
5.4.4 Introgression between the benthic & pelagic ecomorphs	105

	5
5.4.5 Regions of putative adaptive introgression between ecomorphs.....	106
5.4.6 Adaptive introgression candidate genes	107
5.5 Discussion	109
5.5.1 Gene functions under selection & adaptive role	109
5.5.2 Relationship between positive selection and introgression.....	110
5.5.3 Potential adaptive introgression targets in ecological speciation of postglacial fishes.....	111
Chapter 6: General discussion	112
6.1 Summary of thesis chapters.....	112
6.2 Discussion of the general relevance of results.....	114
6.3 Outlook and future directions.....	115
Appendices	117
Appendix A: Figures	118
Appendix B: Tables.....	129
Appendix C: Additional manuscript under review.....	153
Appendix D: Additional published manuscripts	156
References	158

List of Tables

Table 2.1 – Sources of transcriptome assemblies	26
Table 4.1 – Shared outlier genes	83
Table 4.2 – Top 10 overrepresented gene ontology (GO) terms in shared selection outlier regions	87
Table 5.1 – Shifts in introgression in regions under positive selection	106
Table 5.2 – List of adaptive introgression candidate genes.....	107
Table S2.1 – Over- and underrepresented terms from a GO enrichment analysis.	129
Table S2.2 – Orthogroups (OGs) with signals of relaxed and diversifying selection or intensified and diversifying selection.....	135
Table S3.1 – Overview of varied parameters in simulated scenarios	139
Table S4.1 – <i>Grid</i> genes of Arctic charr, human, and zebrafish	140
Table S4.2 – Sampling and coverage information.....	141
Table S4.3 – Admixture log-likelihoods (NGSadmix).....	141
Table S4.4 – Four-taxon test results for lochs Rannoch and Tay	143
Table S4.5 – Metric principal component loadings and explained variance	145
Table S4.6 – Correlation tests between metric pairs.	146
Table S4.7 – Overlap of outlier genes with QTLs	148
Table S5.1 – Introgressed genes under selection in benthic charr	149
Table S5.2 – Introgressed genes under selection in pelagic charr	151
Table S5.3 – Adaptive introgression candidate genes.....	152

List of Figures

Figure 2.1 – Maximum-likelihood phylogenetic tree of salmonids	26
Figure 2.2 – Histogram of the distribution of selection parameter k values.....	32
Figure 2.3 – Selection parameter k distributions for gene ontology (GO) terms	33
Figure 2.4 – Gene Ontology (GO) terms with trends of overrepresentation	34
Figure 2.5 – Top ten enriched biological process GO terms and KEGG pathways	36
Figure 2.6 – Overlapping orthogroups under selection.....	38
Figure 3.1 – General outline of simulated scenarios.....	51
Figure 3.2 – Traces of summary statistics.....	59
Figure 3.3 – Heat map of prediction accuracies for the standard recombination rate region and equal population sizes	60
Figure 3.4 – Suggested steps for detecting the genomic position of selective sweeps ...	61
Figure 4.1 – Sampling locations of Arctic charr.....	74
Figure 4.2 – PCAnsd population structure plots.....	81
Figure 4.3 – Individual genome-wide admixture patterns	82
Figure 4.4 – Allele frequency difference (AFD) between benthic and pelagic ecomorphs	86
Figure 5.1 – Sampling locations of Arctic charr (Dùghaill and Uaine)	98
Figure 5.2 – Schematics of two introgression patterns tested in a four-taxon test	101
Figure 5.3 – SNPRelate population structure plot	103
Figure 5.4 – Manhattan plots of the selection statistic μ and the introgression statistic D across the genome	104
Figure 5.5 – Top 10 overrepresented gene categories in genes under positive selection	105
Figure 5.6 – Top 10 overrepresented gene categories in genes under positive selection and subject to introgression between ecomorphs	108
Figure S3.1 – Heat map of prediction accuracies for the low recombination rate region and equal population sizes	118
Figure S3.2 – Heat map of prediction accuracies for the low recombination rate region and unequal population sizes	119
Figure S3.3 – Heat map of prediction accuracies for the standard recombination rate region and unequal population sizes	120
Figure S3.4 – Heat map of prediction accuracies for the standard recombination rate region and equal population sizes for the SNP window-based approach.....	121
Figure S3.5 – Heat map of prediction accuracies for the standard recombination rate region and unequal population sizes for the SNP window-based approach.....	122
Figure S3.6 – Heat map of prediction accuracies of the location of selective sweeps for the low recombination rate region and equal population sizes for the SNP window- based approach.....	123

Figure S3.7 – Heat map of prediction accuracies for the low recombination rate region and unequal population sizes for the SNP window-based approach	124
Figure S4.1 – Individual genome-wide admixture patterns ($K = 11$ and $K = 16$)	124
Figure S4.2 – Phylogenetic tree of <i>grid</i> genes in Arctic charr, human, and zebrafish....	126
Figure S4.3 – Relationships and correlations between metric pairs.....	127
Figure S4.4 – Allele frequency difference (AFD) between benthic and pelagic ecomorphs	128

Preface

Chapter 2. For this chapter, Kevin Schneider was the sole responsible for the processing of data, analyses, and writing. A version of this chapter has been published in *BMC Genomics*, 20(1), 1-23.

Chapter 3. This chapter was a collaboration with other researchers from the Institute of Biodiversity, Animal Health and Comparative Medicine. Kevin Schneider performed the computer simulations, wrote the computer scripts, performed the analyses, and was responsible for writing. Tom J. White wrote additional computer scripts for analysis and, together with Sonia Mitchell and Richard Reeve, contributed a novel analysis method. Richard Reeve contributed to the method explanations in the manuscript. A version of this chapter has been published in the *Journal of Evolutionary Biology*, DOI: <https://doi.org/10.1111/jeb.13738>.

Chapter 4. For this project, Kevin Schneider performed the lab work, analysed the data, and wrote the manuscript. Arne Jacobs, Oliver E. Hooker, Madeleine Carruthers, and Colin E. Adams provided samples. This chapter is close to submission to *Nature Ecology and Evolution*.

Chapter 5. For this chapter, Kevin Schneider performed the lab work, analysed the data, and wrote the manuscript. Oliver E. Hooker and Colin E. Adams provided samples. This chapter will be submitted to *Molecular Ecology*.

All chapters were supervised by Kathryn R. Elmer and Colin E. Adams

Acknowledgements

I want to express my gratitude to my supervisors, Kathryn R. Elmer and Colin E. Adams, for their unconditional support and help in all matters. Kathryn R. Elmer has shaped my scientific development and progress beyond the lab, has given advice, and has offered her help even in unfavourable conditions. Colin E. Adams has lent his advice and insights (and samples) from all over the salmonid world and was always there when big biological questions required his knowledge and expertise, while always trying to help in other matters.

All lab members have made my experience at the University of Glasgow so much more meaningful and humorous. From the very beginning of my PhD, Arne Jacobs, Hans Recknagel, and Marco Crotti have given advice, enlivened scientific discussions, and provided banter of all kinds. Madeleine Carruthers, James D. Burgon, Elizabeth Kilbride, and Aileen Adam have been excellent advisors on local Scottish peculiarities. Maria Capstick and J. Peter Koene have enriched morning coffee sessions at a previously unseen level. All of you roof lab folks have been great! In fact, this applies to the whole institute.

The most important supporters in my life have always been my parents. Not once have they questioned the paths I have chosen. Their unlimited, stoic, and unconditional help, their kindness and their life advice are the roots of the person I am. My brother has been a true friend, a lantern in difficult times and a spring of joy in good ones. My wife Danyang joined me in the first few months after my arrival in the UK and has never let go. Her love has been like a lucky charm, always by my side.

Author's declaration

I declare that this thesis, and the work contained within it, is my own work unless otherwise stated. No part of this thesis has been submitted towards the fulfilment of any other degree. The work was carried out between October 2016 and November 2020 at the University of Glasgow under the supervision of Professor Kathryn R. Elmer and Professor Colin E. Adams.

Kevin Schneider

November 2020

Signature:

Chapter 1: General introduction

From the very beginning of evolutionary biology in the 19th century, the core of this rapidly developing scientific discipline has been the study of the origin of biological diversity (Darwin 1859; Wallace 1871). For a long time, the understanding of the origins of diversity were based on phenotypic observations and comparative studies of particularly diverse taxa and studies of how external geographic, geological, climatic, and ecological factors shaped this diversity (Weismann 1868, 1872). With the rise of genetics and the appearance of population genetics in the first half of the 20th century, another pressing question emerged, the question of how and to what extent phenotypic diversity is encoded in the genetic material of organisms and how it changes under the influence of evolutionary processes (Haldane 1924; Fisher 1930). However, traditional genetic methods that focussed on only a few, often neutral genetic loci were unable to conclusively answer this question and were mostly reduced to inferring phylogenetic relationships of populations and taxa, calculating genetic diversity, and estimating gene flow among populations (Kimura and Ohta 1971; Nei 1973; Friedman and Adams 1985; Graur 1985; Lowenstein 1986). Natural selection, the crucial shaping force of evolution, could only be studied on a gene-by-gene basis, which led to hit-or-miss scenarios that were largely dependent on the genes being studied (Kimura 1961; Czelusniak et al. 1982; Lawlor et al. 1990).

The advent of so-called next-generation sequencing techniques revolutionised the research on the genetic basis of biological diversity and evolutionary genetics by shifting the field towards a genome-wide view of evolution (Koonin and Mushegian 1996; Eyre-Walker 1999). Even more recently, whole-genome resequencing and whole-transcriptome sequencing have come into reach for an ever-growing number of species (Hudson 2008). The arguably most revolutionary aspect of these new technologies and the amount of data they produce is the possibility to simultaneously analyse the whole set or at least a majority of genes, be it coding or non-coding loci, by studying signals of evolutionary forces, such as genetic drift, gene flow, mutation, recombination, and natural selection (Bonin 2008). Whole-genome techniques have therefore finally made it possible to observe the traces of natural selection, letter by letter.

Despite all the enthusiasm after the arrival of evolutionary genomics, several constraints have made it difficult to gain biological insights from genomic data. First of all, whole-genome data is intrinsically complex and difficult to analyse with conventional population genetic methods. This has led to a time-lag between the availability of large-

scale genomic data and the availability of appropriate analytical methods for the study of natural selection, introgression, and demographic inference (Schadt et al. 2010). Second, a large proportion of biological systems are inappropriate for the systematic study of the genomic origins and underpinnings of biological diversity. This has several reasons, among which are 1) often challenging and intractable population structure of species as well as non-independence of populations (Vitti et al. 2013), 2) relatively slow current phenotypic and often also genomic evolution as a consequence of most adaptations having evolved millions of years ago, which makes inferring population genomic signals more difficult (Racimo 2016; Peyrégne et al. 2017), and 3) the scarcity of repeated evolution of the same characteristics in the same taxa, which makes it difficult to explicitly test the predictability of evolution by examining replicability of adaptations (Lobkovsky and Koonin 2012; Bolnick et al. 2018).

Because it accommodates these challenges, the salmonid genus *Salvelinus* and in particular the circumpolar species Arctic charr (*Salvelinus alpinus*) have become increasingly popular in the study of evolutionary genomics of diversity (Elmer 2016). Arctic charr is characterised by repeated cases of rapid phenotypic diversification in freshwater lakes of the northern hemisphere, which are thought to have originated, at least in part, due to independent colonisation, divergence, and diversification events (Kristjánsson et al. 2012; Elmer 2016; Jacobs et al. 2020). Moreover, the diversification observed in Arctic charr is in many cases parallel or convergent, with divergence in lakes happening mostly along the benthic-pelagic axis into locally adapted, sympatric ecomorphs (Adams et al. 2007; Elmer 2016; Jacobs et al. 2020). Lastly, similar parallel, repeated diversifications also occurred in another salmonid genus, whitefish (*Coregonus*) (Pigeon et al. 1997; Østbye et al. 2006). This makes it possible to screen for genetic similarities between these two salmonid genera in order to obtain genes of putative importance for rapid diversifications.

1.1 The origins of biological diversity

It has long been hypothesised whether and to what extent biological diversity is a consequence of ecological conditions, genomic properties, or both (Yoder et al. 2010). That changing ecological conditions also change the fitness landscapes of populations and species and consequently drive the evolution of phenotypic traits has been found to have both a strong theoretical and empirical basis (Mahler et al. 2010; Yoder et al. 2010). However, the observation of taxa with seemingly higher potential for diversification has kindled an interest in potential genetic characteristics enabling more rapid adaptation to

local environments (Salzburger 2018; Naciri and Linder 2020). This research on the genetic underpinnings of diversification has in many cases been inconclusive, and the search for "speciation genes" has been deemed unrealistic or overly optimistic, as many rapidly diversifying or radiating taxa only diversify under some ecological conditions but not under others (Yoder et al. 2010). This has led to the insight that both specific ecological and genomic conditions need to be fulfilled in order for taxa to diversify rapidly (Yoder et al. 2010).

It is therefore likely the interplay between ecology and genetics that has led to such awe-inspiring diversifications such as the radiation of cichlids in East African and Central American lakes (Brawand et al. 2014; Elmer et al. 2014; Salzburger 2018), the repeated ecological adaptations of *Anolis* lizards on Caribbean islands (Mahler et al. 2013), and the rapid, repeated, parallel, and pervasive adaptive diversifications of salmonids such as whitefish (*Coregonus*) (Pigeon et al. 1997; Østbye et al. 2006) and charr (*Salvelinus*) (Elmer 2016; Jacobs et al. 2020). Ecological opportunity is crucial in several ways. First, strong competition in newly colonised environments can hinder diversification or even lead to local extinction of taxa (Case and Taper 2000; Rabosky 2013). Second, environments must provide a certain set of external conditions to make it possible for ecological niches to be filled (Yoder et al. 2010). Third, environmental conditions must enable population sizes to increase to a certain level in order for natural selection to unleash its full potential in the face of the adverse effects of genetic drift (Whitlock 2000). Ecological opportunity can therefore be considered a crucial prerequisite for diversifications to happen.

Nevertheless, while ecological opportunity can enable diversification in some taxa, it does not to the same degree in other taxa (Yoder et al. 2010). How rapidly diversifying taxa differ from barely or slowly diversifying taxa has thus been an important focus of evolutionary research. In the last few years, diversification has been attributed to properties such as the mutation rate enabling more rapid emergence of adaptive alleles (Lynch et al. 2016; Swings et al. 2017), the occurrence of inversions and other genomic properties reducing the locally destructive effects of recombination (Fan and Meyer 2014), phenotypic plasticity and plastic gene expression regulation as a basis for adaptation and survival in newly colonised habitats (Pfennig et al. 2010; Muschick et al. 2011; Sikkink and Snell-Rood 2016), genetic diversity and the size of the pool of standing genetic variation, which can provide the raw material for rapid diversification (Feulner et al. 2013; Marques et al. 2018; Urban et al. 2020), specific ecological characteristics of species that allow more effective colonisation of new habitats (Schluter 1996), sexual selection as a driver of the build-up of reproductive barriers (Wagner et al. 2012), and the coupling of

genetic variants with specific barrier loci, which, in turn, decrease the local recombination rate relative to the genomic background (Ravinet et al. 2017).

1.2 Evolutionary research based on genome-scale data

Reduced-representation sequencing techniques have been one of the most popular approaches for obtaining genome-wide data, due to their cost efficiency and relatively easy application to non-model organisms, as they do not require a reference genome and can be applied using earlier sequencing platforms that are more widely available (Davey and Blaxter 2010). However, there is mounting evidence that the utility of RAD-sequencing for inferring genome-wide patterns of selection has its limits (Lowry et al. 2017). In the last few years, due to advances in the availability of high-throughput sequencing machines and a strong reduction of sequencing costs, sequencing whole-genomes has become more practical and feasible, which has led to significant progress in the study of ecological diversifications at the whole-genome level (Jones et al. 2012; Malinsky et al. 2018; Meier et al. 2018; Nosil et al. 2018; Morales et al. 2019). This has been in part also due to advances in third-generation sequencing techniques and molecular genetic methods that facilitated the assembly of draft reference genomes, even for research projects and groups that some years ago would have lacked the resources for the independent assembly of draft genomes (Bleidorn 2016). Moreover, combined with a reduction of reagent volumes and less expensive and more efficient library preparation methods, low-coverage sequencing of whole genomes has opened up a range of possibilities for population genomic studies that had been unthinkable just a decade ago (Therkildsen and Palumbi 2017). Importantly, one of the main insights has been that answering population genomic questions does not necessarily require deep sequencing of whole genomes but can also be based on relatively low efforts of sequencing that likely miss some genomic variants and/or are accompanied by higher uncertainty of variant assignment (Buerkle et al. 2013; Rustagi et al. 2017). One of the reasons for the increasing popularity of low-coverage sequencing is that population genomics has not only focussed on the discovery of specific variants at a nucleotide position but also on signals of evolutionary processes at a genomic region level (Ellegren 2014).

1.3 Methods for analysing genome-scale data and population genetic signatures

There has been significant progress in using this kind of low-coverage whole-genome resequencing data in population genomic studies of ecological model and non-model organisms, partly owing to the use of genotype likelihoods instead of relying on SNP or genotype calling (Nielsen et al. 2011). Many methods commonly used in population genomic studies are now incorporated in the genotype likelihood-based software framework ANGSD (analysis of next generation sequencing data) (Korneliussen et al. 2014), which allows for the determination of population structure, the performance of tests for introgression, calculation of population genetic summary statistics, and genome-wide association analyses, amongst others. Such or similar approaches based on genotype likelihoods have been applied to various systems of ecological and evolutionary genomics, such as Atlantic silversides (*Menidia menidia*) (Therkildsen and Palumbi 2017; Wilder et al. 2020), capuchinos (*Sporophila* spp.) (Campagna et al. 2017), and flycatchers (*Ficedula*) (Burri et al. 2015). Methods that are based on population allele frequencies are suitable for application to low-coverage data as well, for example the F_{ST} alternative AFD (allele frequency difference) (Berner 2019).

In the past decade, analyses of whole-transcriptome data have become more feasible in a similar way, not only because of reduced sequencing costs (Ekblom and Galindo 2011; Huang et al. 2016). In a multispecies comparative context, the inference of orthologous genes or groups of orthologues is essential for being able to infer any kind of patterns of molecular evolution or perform nucleotide variant analysis. Therefore, the development of fast and reliable orthology inference tools, such as OrthoFinder (Emms and Kelly 2015, 2019), OMA (Altenhoff et al. 2010), and eggNOG (Jensen et al. 2007; Huerta-Cepas et al. 2019) has been a prerequisite for associated discoveries on the evolution of protein-coding genes. Similarly, methods for inferring complex evolutionary processes such as positive or diversifying selection, relaxed selection, and intensified selection have been released that can cope with large numbers of taxa as well as large numbers of transcripts or orthologues (Pond and Muse 2005; Smith et al. 2015; Wertheim et al. 2015). Taken together, the ready availability of high-throughput sequencing data and the necessary tools to analyse such kind of data have been fruitful for the study of genome- and transcriptome-scale data in the context of evolutionary genomics and molecular evolution.

1.4 Rapid diversifications as models for the study of diversity evolution

Equipped with the large-scale data and appropriate tools mentioned above, studying rapid ecological diversifications in nature allows us 1) to learn what fundamental processes lead to the formation of species, 2) to draw conclusions about what genomic factors are important in the adaptation of phenotypes to new environments, and 3) to infer the functions and processes that are under the strongest selection pressures, which ultimately tells us something about the biology of the studied organisms (Brawand et al. 2014; Elmer 2016; Salzburger 2018; McGee et al. 2020). In the case of repeated, parallel diversifications, we can also make inferences about what specific genetic elements, genomic regions, and nucleotide changes are subject to the same selective forces in different populations while 'ignoring' such signals that only appear in single populations and therefore likely lack generality for the adaptive process (Hohenlohe and Magalhaes 2020). This approach of screening for sharing of evolutionary signals, such as selection and introgression, also has the advantage that demographic history and stochasticity are less likely to affect several populations in the same way, leaving those signatures that are of potential importance to diversification. Common study systems for rapid diversification have been taxa such as cichlids, which show tremendous radiations from the Great Lakes of East Africa to small crater lakes in Africa as well as Central America (Brawand et al. 2014; Elmer et al. 2014; Malinsky et al. 2015; McGee et al. 2020), *Anolis* lizards, which have diversified into nearly 150 often uniquely adapted species in the Caribbean from putatively only two initial colonising species (Jackman et al. 1997; Mahler et al. 2013), *Timema* stick insects, a species group that repeatedly adapted specific colourations to confer camouflage on their host plants (Nosil et al. 2018), and salmonids, which contain two genera, *Coregonus* (whitefish) and *Salvelinus* (charr), that are characterised by unmatched, rapid, parallel diversifications into ecomorphs adapted to benthic, pelagic, piscivorous, insectivorous, and other types of feeding (Pigeon et al. 1997; Østbye et al. 2006; Adams et al. 2007; Elmer 2016; Jacobs et al. 2020).

1.5 Salmonids and Arctic charr as models for ecological and evolutionary genomics

Salmonids are an economically important group of anadromous and freshwater taxa that evolved remarkable adaptations to their environments, such as unique forms of anadromy and homing behaviour (Keefer and Caudill 2014). However, it is mostly lacustrine populations of the two genera *Coregonus* and *Salvelinus* that repeatedly diversified into

ecomorphs in freshwater lakes (Pigeon et al. 1997; Østbye et al. 2006; Adams et al. 2007; Elmer 2016). *Coregonus* mainly shows diversifications along the depth or benthic-pelagic axis within lakes (Vonlanthen et al. 2009). Its speed of adaptation and diversification has been exemplified by the observation of a recovery from eutrophication-induced speciation reversal in Lake Constance, possibly facilitated by a large pool of standing genetic variation, which was recruited for recent re-divergence and ecological niche expansion (Jacobs et al. 2019).

Some authors even consider Arctic charr to be the most variable species of vertebrates on earth, owing to its unrivalled ability for rapid morphological and ecological divergence in a large number of populations and lineages across its Holarctic distribution (Klemetsen 2013). One of the most striking examples is a population of Arctic charr in southern Greenland, consisting of six genetically distinct clusters belonging to five distinct ecomorphs (Doenz et al. 2019). However, what makes Arctic charr a remarkable model for studying the predictability and replicability of the evolution of local adaptation, are the high number of repeated divergences into benthic and pelagic ecomorphs that make it possible to study numerous instances of adaptation to common environments (Adams et al. 2007; Elmer 2016; Jacobs et al. 2020).

This high potential for diversification and ecological adaptation has been attributed to either phenotypic plasticity (Alexander and Adams 2004) or underlying genetic diversification (Skúlason et al. 1996), or a combination of these factors (Adams and Huntingford 2004). Evidence from reduced-representation sequencing and gene expression data suggests that at least some of the phenotypic diversity is attributable to genetic changes tied to changes in gene expression, while the amount of genetic parallelism is limited (Jacobs et al. 2020). Combined with the evidence of parallelism in gene regulatory networks, this is a strong indication that the specific genetic changes can but do not have to be parallel in order for gene expression and phenotypic changes to be parallel (Jacobs et al. 2020). While results from threespine stickleback and Arctic charr have shown phenotypic plasticity to exacerbate phenotypic evolutionary change (Adams et al. 2004; Wund et al. 2008; Klemetsen 2010), the combined evidence from multiple diversifications has *de facto* ruled out phenotypic plasticity as the sole factor leading to ecomorphological divergence in Arctic charr (Jacobs et al. 2020), which highly suggests a genetic component of local adaptation.

However, the specific genetic changes underlying diversifications in both Arctic charr and whitefish are still unknown, partly due to the limited resolution of reduced-representation sequencing that was used for most of the past studies of these two salmonid taxa (Gagnaire et al. 2013; Rougeux et al. 2017; Jacobs et al. 2019, 2020). Moreover, it is

unknown how demographic history affects the evolutionary signatures that are used for inferring similarities among taxa and populations, which is particularly important for the complex, varying demographics underlying salmonid populations (Pigeon et al. 1997; Gagnaire et al. 2013; Rougeux et al. 2017; Doenz et al. 2019; Jacobs et al. 2020). The use of both transcriptome sequencing and whole-genome resequencing promises to bring closer a detailed genomic picture of what genetic factors and changes are associated with diversification, both in a comparative context across taxa and at an across-population, intraspecies level.

1.6 Thesis outline

In Chapter 2), I attempt to find the similarities of selection signals and sharing of selection targets in two highly diversifying salmonid genera, *Coregonus* and *Salvelinus*, represented by two species each and relative to non-diversifying taxa, using transcriptome data derived from publicly available RNA-sequencing data sets. I assess the amount of relaxation or intensification of selection as well as the occurrence of positive/diversifying selection of protein-coding genes to find genes that are under selection in both highly diversifying salmonid genera and therefore potentially of higher general relevance for the diversification process. This is followed by a detailed analysis of over- and underrepresentation of gene functions of genes under different types of selection.

Chapter 3) is an analysis of the influence of complex demographic histories, such as the ones found in many salmonid populations, on the inference of selective sweeps using various metrics, which are a necessity for many analyses of selection, especially in the case of the computationally challenging analysis of whole-genome data. I used forward-in-time simulations of chromosomes under sympatric divergence and secondary contact scenarios, varying parameters such as the migration rate, recombination rate, population growth, population size, and type of selective sweep. The metrics tested include methods based on the site frequency spectrum, multilocus genotype frequencies, the genotype distance distribution, linkage disequilibrium, and differentiation between populations. I explore the relative performance of each of these metrics in predicting the occurrence and location of selective sweeps, and give recommendations about the application of such metrics under complex demographic scenarios.

Chapter 4) focusses on the sharing of positive selection on genome regions and genes across five benthic-pelagic ecomorph pairs. The analyses are based on whole-genome resequencing of a total of 92 samples and six populations of Arctic charr. I use a variety of methods to infer selection, analyse overrepresentation of gene functions in

selected regions, and infer the types of substitutions in genes under selection. Moreover, I analyse the overlap of selection outlier genes with ecologically relevant quantitative trait loci of salmonids. This is accompanied by analyses of population structure, admixture, and four-taxon tests to elucidate introgression between the ecomorphs of two populations that have putatively been connected by gene flow.

In Chapter 5), I analyse the interplay of positive selection and introgression in a population of Arctic charr that has experienced a secondary contact to screen for genes putatively subject to adaptive introgression between ecomorphs. The basis for the analyses is whole-genome resequencing of 30 samples of two ecomorphs (one benthic and one pelagic) of Loch Dùghaill and a population of Arctic charr of a neighbouring lake that is closely related to the pelagic ecomorph of Loch Dùghaill. I apply an analysis of positive selection based on the joint signals of the site frequency spectrum, nucleotide diversity, and linkage disequilibrium in genomic windows. Introgression is inferred using window-based four-taxon tests to obtain estimates of gene flow between ecomorphs. Both the outlier genes and overrepresentation of gene functions of regions under selection, introgression, and adaptive introgression are examined.

In Chapter 6), the thesis closes with a summary of the results of the data chapters and a general discussion of the insights gained and conclusions drawn from these results, complemented with a section on the outlook and future directions of ecological genomics and molecular evolution in salmonids.

Chapter 2: Parallel selection on ecologically relevant gene functions in the transcriptomes of highly diversifying salmonids

2.1 Abstract

Salmonid fishes are characterised by a very high level of variation in trophic, ecological, physiological, and life history adaptations. Some salmonid taxa show exceptional potential for fast, within-lake diversification into morphologically and ecologically distinct variants, often in parallel; these are the lake-resident charr and whitefish (several species in the genera *Salvelinus* and *Coregonus*). To identify selection on genes and gene categories associated with such predictable diversifications, we analysed 2,702 orthogroups (4.82 Mbp total; average 4.77 genes/orthogroup; average 1,783 bp/orthogroup). We did so in two charr and two whitefish species compared to five other salmonid lineages, which do not evolve in such ecologically predictable ways, and one non-salmonid outgroup.

All selection analyses are based on *Coregonus* and *Salvelinus* compared to non-diversifying taxa. We found more orthogroups were affected by relaxed selection than intensified selection. Of those, 122 were under significant relaxed selection, with trends of an overrepresentation of serine family amino acid metabolism and transcriptional regulation, and significant enrichment of behaviour-associated gene functions. Seventy-eight orthogroups were under significant intensified selection and were enriched for signalling process and transcriptional regulation gene ontology terms and actin filament and lipid metabolism gene sets. Ninety-two orthogroups were under diversifying/positive selection. These were enriched for signal transduction, transmembrane transport, and pyruvate metabolism gene ontology terms and often contained genes involved in transcriptional regulation and development. Several orthogroups showed signs of multiple types of selection. For example, orthogroups under relaxed and diversifying selection contained genes such as *ap1m2*, involved in immunity and development, and *slc6a8*, playing an important role in muscle and brain creatine uptake. Orthogroups under intensified and diversifying selection were also found, such as genes *syn3*, with a role in neural processes, and *ctsk*, involved in bone remodelling.

Our approach pinpointed relevant genomic targets by distinguishing among different kinds of selection. We found that relaxed, intensified, and diversifying selection affect orthogroups and gene functions of ecological relevance in salmonids. Because they were found consistently and robustly across charr and whitefish and not other salmonid

lineages, we propose these genes have a potential role in the replicated ecological diversifications.

2.2 Introduction

Identifying the molecular mechanisms underlying adaptive phenotypic divergence is a central challenge for evolutionary biology; a key first step is to detect genes under selection rather than reflecting background neutral evolutionary processes. Parallel or convergent evolution at the molecular level may, or may not, be associated with phenotypic parallelisms across species, but the idea remains compelling (Manceau et al. 2010; Elmer and Meyer 2011; Berner and Salzburger 2015) and has been an important analytical framework to advance research in non-model systems (Brawand et al. 2014; Foote et al. 2015; Torres-Dowdall et al. 2015). Molecular parallelism or convergence can be inferred either from nucleotide site-specific changes (Castoe et al. 2009; Dobler et al. 2012; Bailey et al. 2015; Foote et al. 2015) or at a higher level, in the sense of similar genes being targeted by similar selective forces (Farhat et al. 2013; Laayouni et al. 2014; Chikina et al. 2016; Castiglione et al. 2018).

Fishes have proven a powerful ecological and evolutionary group for comparisons of genes under selection and that are associated with ecological and evolutionary novelty. Sticklebacks, for example, have become a model group of repeated ecological adaptation of Holarctic marine and freshwater distribution (Hartfield and Schluter 1999; Berner et al. 2009; Schluter and Conte 2009). In cichlid fishes, adaptive potential and highly malleable phenotypes are spread throughout the family. In some cases, it has been shown that relaxed selection (Houle et al. 1994; Yoder et al. 2010; Hunt et al. 2011; Velotta et al. 2014) or positive (diversifying) selection (Elmer et al. 2010; Fan et al. 2011) correlate with phenotypic diversification. However, ecological opportunity differs dramatically among cichlid lineages (Wagner et al. 2012; Recknagel et al. 2014; Koblmüller et al. 2015; Arbour and López-Fernández 2016; Stroud and Losos 2016), which makes it difficult to pinpoint taxa in which adaptive potential is elevated due to a shared genetic toolset (Berner and Salzburger 2015; Torres-Dowdall et al. 2015). In contrast, freshwater lake-resident salmonids of different species and genera have similar ecological opportunity and commonly sympatric distributions across the northern hemisphere (Matthews 1998; Stearns and Hendry 2004; Griffiths 2006). Furthermore, the freshwater habitats of northern fishes were all colonised on a similar postglacial timescale (Bernatchez and Wilson 1998; Griffiths 2006), unlike the dramatically different and complex colonisation histories of

cichlids (Elmer et al. 2009; López-Fernández et al. 2013; Brawand et al. 2014; Matschiner et al. 2017).

Salmonid fishes are increasingly used as model organisms in evolutionary research, because of their ecological diversity, economic value, and replicated evolution of distinct ecomorphs and traits in some taxa, such as depth specialisation and alternative migratory tactics (Primmer 2011; Dodson et al. 2013; Filteau et al. 2013; Chavarie et al. 2016; Elmer 2016; Laporte et al. 2016; Macqueen et al. 2017; Moore et al. 2017). Two salmonid genera in particular, the whitefishes (*Coregonus*) and the charrs (*Salvelinus*), are not sister taxa but exhibit parallel (or convergent) adaptive tendencies in freshwaters across the northern hemisphere. They have repeatedly diverged into various within-lake ecomorphs along the depth axis over short evolutionary time spans that are unmatched in any other salmonid species (Jonsson and Jonsson 2001; Kottelat and Freyhof 2007; Vonlanthen et al. 2009; Bernatchez et al. 2010; Elmer and Meyer 2011; Muir et al. 2016). The evolutionary and molecular basis for why *Coregonus* and *Salvelinus* show such a high degree of ecomorphological adaptability while other salmonid species do not is, however, unknown (Bernatchez et al. 2010; Elmer 2016; Muir et al. 2016).

Determining how single and combined effects of selection act at the molecular level is facilitated by new analytical tools (Pond and Muse 2005; Wertheim et al. 2015; Smith et al. 2015). These selective processes are associated with adaptive evolution in different ways and are most powerful when linked with known lineage-specific phenotype changes or phenotypic diversification (Lamichhaney et al. 2016; Morandin et al. 2016). Two selective processes – relaxed and intensified selection – are on opposite ends of the spectrum. Relaxed selection decreases the selective constraints of a gene and can lead to the accumulation of nonsynonymous substitutions and consequently changes in the amino acid sequence of a protein. By releasing a gene of selective constraints, relaxed selection can potentially foster phenotypic novelty, plasticity, and evolutionary innovation (Hunt et al. 2011; Brawand et al. 2014; Wertheim et al. 2015; Sharma et al. 2018). In contrast, intensified selection increases selective constraints but can also manifest as positive intensified selection leading to more differences at some sites of a gene (Finseth et al. 2014; Afanasyeva et al. 2018). Intensified selection implies changes to a gene have strong fitness consequences (Wertheim et al. 2015). Additionally, lineage-specific episodic diversifying selection, or positive selection, will leave other signatures at the sequence level, such as more nonsynonymous changes than expected under neutrality at a subset of positions in a gene on some branches in the phylogenetic tree (i.e. branch-site model) (Yang and Nielsen 2002; Smith et al. 2015). While relaxed and intensified selection are

antithetical, in either case diversifying (positive) selection can simultaneously act at a proportion of sites in a gene (Bielawski and Yang 2001; Murrell et al. 2015).

It has long been proposed that the propensity for ecological speciation in some salmonid lineages is associated with shared patterns of relaxed or diversifying selection on ecologically relevant genes and gene function terms (Orr et al. 2004; Benderoth et al. 2006; Elmer and Meyer 2011; Hunt et al. 2011; Berner and Salzburger 2015; Elmer 2016). We focus on the well-characterised and richly diversifying genera *Coregonus* and *Salvelinus*, which show repeated within-lake divergences into distinct ecomorphs across the northern hemisphere (Bernatchez et al. 2010; Muir et al. 2016). The *Coregonus* species assessed are lake resident at least since postglacial times and have high rates of within-lake adaptive divergence (Vonlanthen et al. 2009; Bernatchez et al. 2010; Hudson et al. 2017). *Salvelinus* species are mostly freshwater residents and have undergone frequent adaptive divergence into ecomorphs along the depth axis, predominantly within lakes (Adams et al. 1998; Chavarie et al. 2015, 2016; Muir et al. 2016). Representatives of all other major salmonid lineages, *Oncorhynchus*, *Salmo*, and *Thymallus*, were also included in the dataset; these are generally riverine or anadromous genera that do not extensively diversify within lakes (Klemetsen et al. 2003; McPhee et al. 2007). By assessing consistency across two non-sister lineages, *Coregonus* and *Salvelinus*, our approach mitigates against false positives. The focus on orthogroups within and across species, rather than single genes, alleviates the problem of differing relaxation of selective constraint in duplicated compared to non-duplicated or rediploidised genes (Hodgins et al. 2015), which is particularly important in salmonids due to the whole-genome duplication (WGD) that their common ancestor experienced 80 – 103 Mya (Macqueen and Johnston 2014; Lien et al. 2016; Robertson et al. 2017).

Here, we use a genome-wide comparative approach to test for shared evidence of selection at particular categories of genes, gene functions, and gene ontology terms in the two highly diversifying lineages, *Coregonus* and *Salvelinus*, relative to all other major salmonid lineages. We test a comprehensive suite of 2,702 orthologous protein-coding gene sets (orthogroups) for signals of parallel relaxed, intensified, and diversifying/positive selection in *Coregonus* and *Salvelinus* (average of 4.77 genes per orthogroup; 4.82 Mbp in total; average of 1,783 aligned bp per orthogroup). By distinguishing among different kinds of selection in replicate across two independent lineages, our approach can pinpoint the action of selective pressure more accurately. We find that different types of selection target different gene sets and functions in salmonids, with novel and established ecological relevance for repeated, parallel diversification potential.

2.3 Materials & methods

2.3.1 Taxonomic dataset

All analyses were performed on RNAseq transcriptome datasets drawn from a search in NCBI and a literature search in Google Scholar (keywords: each salmonid genus and species, in combination with "transcriptome" and/or "RNAseq" or "RNA-seq") to identify all available studies. The final dataset represents one transcriptome per species of transcriptomes of all available taxa as of October 2016, with additional in-house data for *Salmo salar*, *Salmo trutta*, *Salvelinus alpinus*, and *Coregonus lavaretus* that was published at a later point (Carruthers et al. 2018). We aimed to retain transcriptomes with a maximal number of overlapping transcripts across salmonids, which consisted of nine species representing all five major genera (*Coregonus*, *Oncorhynchus*, *Salmo*, *Salvelinus*, and *Thymallus*), and the closest relative as outgroup (*Esox lucius*) (Figure 2.1, Table 2.1). The selected transcriptome assemblies were all derived from several tissues so as to obtain almost complete sets of transcripts (see references in Table 2.1). The topology for foreground branch definitions is based on the phylogenetic tree shown in Figure 2.1, which was obtained using maximum-likelihood tree estimation based on the 1,564 1-to-1 orthologues present in the data (derived from the same set of transcriptomes; 1-to-1 orthologue alignments available upon request) and in agreement with the current understanding of salmonid phylogenetic relationships (Macqueen et al. 2017; Lecaudey et al. 2018). From the full dataset, ranging from 59.0 to 209.2 Mbp per transcriptome (Table 2.1), we obtained 2,702 sequences of orthogroups (mean number of orthologues per orthogroup: 4.77; mean length 1,783 bp; range 607–14,743 bp; total 4.82 Mbp). Using a *Blast2GO* (Conesa et al. 2005) analysis (for details see below), we confirmed that the dataset was not enriched for any particular GO categories (Fisher's Exact Test: all $P > 0.10$) and therefore not a biased representation of transcripts.

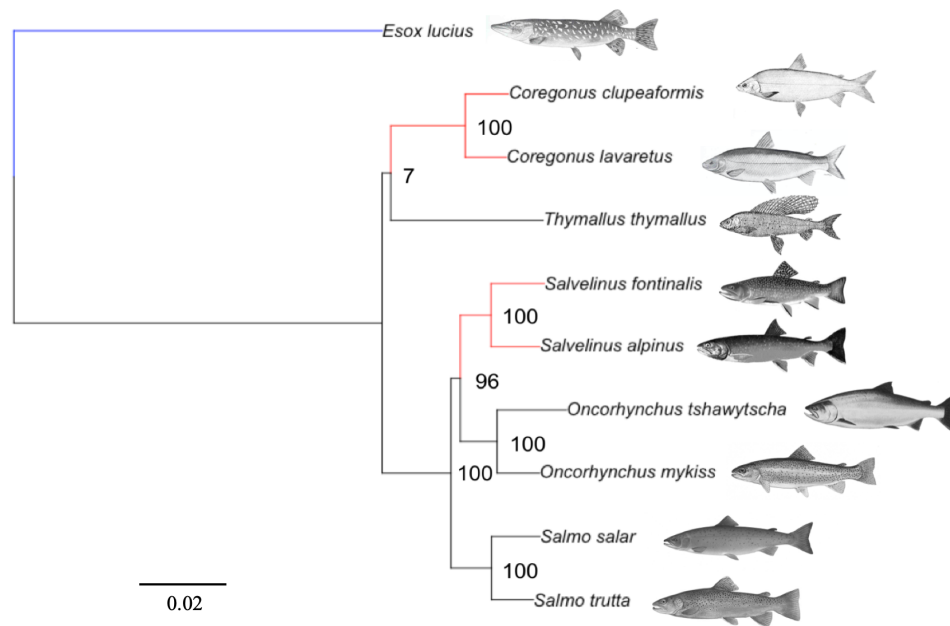


Figure 2.1 – Maximum-likelihood phylogenetic tree of nine salmonid species (foreground (diversifying) taxa *Coregonus* and *Salvelinus*: red branches) and northern pike (*Esox lucius*: blue branch; used as outgroup). Node support values are bootstrap values from 1000 bootstrap replications. *Coregonus* and *Salvelinus* are two genera with exceptional ability for repeated, rapid diversification into ecomorphs within lakes that is unmatched in other salmonid taxa (Jonsson and Jonsson 2001; Kottelat and Freyhof 2007; Vonlanthen et al. 2009; Bernatchez et al. 2010; Elmer and Meyer 2011; Muir et al. 2016). The phylogeny was calculated based on the 1,564 1-to-1 orthologues present in the data (derived from the same set of transcriptomes; 1-to-1 orthologue alignments available upon request). Branch lengths correspond to the number of substitutions per site. The scale bar indicates the average number of substitutions per nucleotide site. All pictures used here are under public domain.

Table 2.1 – Sources of transcriptome assemblies used in this study. Transcr. = Transcript

Transcriptomes				
<i>Species</i>	<i>Source/Accession no.</i>	<i>Size</i>	<i>Transcr.</i>	<i>Reference</i>
		<i>[Mbp]</i>	<i>count</i>	
<i>Coregonus clupeaformis</i> (lake whitefish)	PhyloFish (phylofish.sigenae.org, last access: 05.10.2016)	158.5	66,996	Pasquier et al. 2016
<i>Esox lucius</i> (northern pike)	GATF00000000.1 (NCBI GenBank)	188.7	55,424	Rondeau et al. 2014
<i>Oncorhynchus mykiss</i> (rainbow trout)	Salmon Transcriptome website (salmon.cgrb.oregonstate.edu, last access: 05.10.2016)	209.2	130,599	Fox et al. 2014

<i>Oncorhynchus tshawytscha</i> (Chinook salmon)	GSE59756 (NCBI Gene Expression Omnibus)	159.8	183,740	Tomalty et al. 2015
<i>Salvelinus fontinalis</i> (brook trout)	PhyloFish (phylofish.sigenae.org, last access: 05.10.2016)	166.0	69,441	Pasquier et al. 2016
<i>Thymallus thymallus</i> (European grayling)	PhyloFish (phylofish.sigenae.org, last access: 05.10.2016)	160.9	67,157	Pasquier et al. 2016
<i>Coregonus lavaretus</i> (European whitefish)	earlier version of GGAO000000000.1 (NCBI GenBank; made accessible upon request)	59.0	44,348	Carruthers et al. 2018
<i>Salvelinus alpinus</i> (Arctic charr)	earlier version of GGAP000000000.1 (NCBI GenBank; made accessible upon request)	59.5	44,354	Carruthers et al. 2018
<i>Salmo salar</i> (Atlantic salmon)	earlier version of GGAQ000000000.1 (NCBI GenBank; made accessible upon request)	63.0	44,554	Carruthers et al. 2018
<i>Salmo trutta</i> (brown trout)	earlier version of GGAR000000000.1 (NCBI GenBank; made accessible upon request)	62.0	46,308	Carruthers et al. 2018
Genome				
<i>Species</i>	<i>Source/Accession no.</i>	<i>Size [Mbp]</i>	<i>Gene count</i>	<i>Reference</i>
<i>Salmo salar</i> (Atlantic salmon)	GCF_000233375.1 (NCBI RefSeq assembly)	77.5	48,589	Lien et al. 2016

2.3.2 Orthogroup and OG tree inference

The longest open reading frames (ORFs) in the transcriptome sequences of the ten species (excluding mitochondrial genes due to different genetic codes that could bias the selection analyses) and, additionally, protein-coding genes of the salmon genome (NCBI RefSeq assembly accession: GCF_000233375.1) (Lien et al. 2016) were determined using *TransDecoder* v2.1 (<https://github.com/TransDecoder/TransDecoder>). Next, the standard mode of *OrthoFinder* v2.2.3 (Emms and Kelly 2015), which uses a five-step algorithm that mitigates gene length bias and other biases introduced by other methods, was used to infer orthogroups from the *TransDecoder* ORFs. Based on the *OrthoFinder* orthogroup assignments, gene sequences were then extracted from the whole transcriptomes with custom *Python* scripts (*Python* v3.6.5) if they were also present in the protein-coding sequences of the Atlantic salmon genome v2 (NCBI RefSeq assembly accession:

GCF_000233375.1) (Lien et al. 2016) to exclude spurious transcripts. To further minimise the number of spurious transcripts retained, only orthogroups with a higher or equal number (but not larger than four) of orthologous genes in the focal species compared to the salmon genome were included to avoid inclusion of collapsed orthologues (i.e., multiple genes that were merged into one during assembly, which seems to be a common issue in the salmonid transcriptomes published so far according to in-house results and A. Yurchenko and M. Carruthers, personal communication). Orthologues were then combined into orthogroups using custom *Python* scripts if they matched the same gene sets in the salmon reference genome. Duplicated orthogroups, inferred as multiple occurrences of the same orthologue sets, orthogroups with a number of sequences exceeding 80, which could include an inflated number of duplicates, or lower than 7, to avoid orthogroups unrepresentative of the whole dataset, and those with fewer than 7 taxa were removed with custom *Python* scripts. The longest ORFs with a minimum length of 200 bp were then extracted from this orthogroup set with `get_orfs_or_cdss.py` (https://github.com/peterjc/pico_galaxy/blob/master/tools/get_orfs_or_cdss/get_orfs_or_cdss.py). A codon-based alignment was then produced using *PRANK* v.140603 [158]. The aligned orthogroups were then trimmed and filtered with *trimAl* (-resoverlap 1.0, -seqoverlap 0.38 -noallgaps) (Capella-Gutiérrez et al. 2009), optimising both alignment quality and the number of retained alignments. Using the final set of orthogroups as input for *FastME* (Lefort et al. 2015), we then constructed phylogenies of each orthogroup separately for use in the selection analyses below. The number of retained genes per orthogroup was calculated using custom *Bash* (*Bash* v3.2) and *Python* scripts.

2.3.3 Selection analyses

Foreground branches for subsequent selection analyses were labelled using a custom *R* script (*R* v3.5.1) and the *R* package *ape* (Paradis et al. 2004; R Core Team 2016; Popescu et al. 2012). All selection analyses were then performed in *HYPHY* v2.3.13.20180521beta(MP) (Pond and Muse 2005). Orthogroup alignment FASTA file headers were shortened using custom *Python* scripts and FASTA and tree files were combined with custom *Bash* scripts. Using the above foreground definitions, *RELAX* (Wertheim et al. 2015), included in the *HYPHY* package, was run to infer orthogroups with significant signals of relaxed or intensified selection in the foreground branches relative to the background branches and to infer the selection strength parameter k (as defined in Wertheim et al. 2015), which can be thought of as the variance in positive selection of a sequence and ranges from 0 to 50, where values < 1 indicate relaxed selection and values $>$

1 indicate intensified selection. The software *aBSREL* (adaptive Branch-Site Random-Effects Likelihood inference) (Smith et al. 2015), also contained in the *HYPHY* package, was then used to obtain orthogroups with a proportion of sites under significant diversifying or positive selection in the foreground branches relative to the background branches. Relevant information from both the *RELAX* and *aBSREL* output was extracted with custom *Bash* scripts. Multiple testing was accounted for with false-discovery rate (FDR) correction (Benjamini and Hochberg 1995) or Bonferroni correction (Bonferroni 1936; Dunn 1961) in *R*. We report on the FDR level of significance throughout the manuscript. Significant differences in the number of orthogroups under relaxed or intensified selection were quantified using Fisher's Exact Tests in *R*. A two-sided one-sample Wilcoxon signed-rank test in *R* was used to infer whether the overall selection parameter (k) distribution differed significantly from the null expectation of 1. Selection parameter distributions were visualised using the *R* package *ggplot2* (Wickham 2011). Hypergeometric expectations of orthogroups under multiple types of selection were calculated in *R*.

2.3.4 Gene ontology analyses

GO terms were inferred using a complementary set of methods in *Blast2GO* v5.2.5 (Conesa et al. 2005). We used *BLASTn* (*megablast*) (Altschul et al. 1990; Morgulis et al. 2008) using *QBLAST* (NCBI Blast Server) and standard web blast against non-redundant nucleotide sequences and in a second step against protein sequences. A BLAST expectation value (e -value) threshold of 1.0E-3 was used. Only the top blast hit was retained per orthogroup blast search. Blast descriptors were extracted for later annotations. *BLASTn* was run with a word size of 28, a low complexity filter, a High Scoring Pair (HSP) length cutoff of 33, and an HSP-hit coverage of 0. GO mapping was performed with the *Goa version 2018.12* database (Camon et al. 2004). Orthogroups were remapped with *InterProScan* (Zdobnov and Apweiler 2001; Quevillon et al. 2005; Jones et al. 2014) using *EMBL-EBI InterPro* to improve mapping success. Annotations were then created using an annotation cut-off of 55, GO weight of 5, an e -value filter of 1.0E-6, an HSP-hit coverage cut-off of 0, and a hit filter of 500. Further annotations were obtained using merging of *InterProScan* GOs with existing annotations. Annotations were then augmented by *ANNEX* (Myhre et al. 2006). Remaining unannotated sequences were annotated using blast descriptions, applying a minimum similarity of 85 and validation of annotations. Gene functions were inferred based on *Blast2GO* and Uniprot/Swiss-Prot gene information for fish, mice, and humans (last accessed on 27/01/2019).

GO term enrichment analyses were performed using Fisher's Exact Tests in *Blast2GO* on the whole dataset and all sets of orthogroups under selection (from *RELAX* and *aBSREL* with Bonferroni multiple correction, FDR correction, or no correction) with FDR multiple correction (FDR threshold of 0.10). GO terms were extracted from the *Blast2GO* output using custom *R* scripts. Overlap of sets of orthogroups under selection was determined with custom *R* scripts. Additionally, GO term enrichment was tested using custom *R* scripts to see trends in over- and underrepresentation (*Blast2GO* only outputs significantly over- and underrepresented GO terms after FDR correction). Violin plots of selection parameter k distributions, compared across GO terms, were plotted with the *ggplot2* *R* package. The significance of deviations from the null expectation of $k = 1$ was tested for each GO term using two-sided one-sample Wilcoxon signed-rank tests in *R*.

To summarise and visualise the obtained GO terms in the different sets of orthogroups under selection, we used the *REVIGO* online analysis tool with an allowed similarity score of GO terms of 0.9, the zebrafish reference GO term database available in *REVIGO*, and SimRel as semantic similarity measure (last access: 01.10.2019) (Supek et al. 2011), which makes use of a semantic clustering algorithm and the P values from the GO term enrichment analysis in *R* above. In this context, semantic similarity refers to the degree of relatedness between two entities/GO terms by the similarity in meaning of their annotations (Pesquita et al. 2009). *REVIGO* clusters GO terms of sets of genes based on this algorithm in a two-dimensional semantic similarity space. Additionally, *REVIGO* ranks the GO terms of genes according to their redundancy.

2.3.5 Gene set enrichment analysis

Gene set enrichment analysis (GSEA) was performed on gene symbols of zebrafish homologues of human genes derived from *Blast2GO*. Zebrafish homologues were obtained from http://www.informatics.jax.org/downloads/reports/HOM_AllOrganism.rpt using a custom *R* script. The selection parameter (k) values from the above *RELAX* analysis were transformed into -1 to 1 scores in *R*, with negative values representing relaxed selection orthogroups and positive values representing intensified selection orthogroups. We then used *WebGestalt* (Zhang et al. 2005; Wang et al. 2013, 2017; Liao et al. 2019) on the zebrafish gene symbols and associated transformed k scores to derive a) the top ten enriched biological process GO terms and b) the top ten enriched KEGG pathways in both the relaxed and intensified orthogroups. We performed 10,000 permutations for each GSEA run. All other parameters were kept at default settings (last use: 02.10.2019).

2.4 Results

2.4.1 Selection parameter distribution and number of orthogroups under selection

Shared molecular response to selection in two whitefish and two charr species was inferred relative to six background species (five salmonids and one pike, Figure 2.1). The selection parameter k in whitefish and charr, ranging from 0 (very relaxed) to 50 (very intensified), had a median value of 0.992 across orthogroups and was significantly different from the neutral expectation of 1 (Wilcoxon signed-rank test: $V = 1,995,900$, $P = 8.859\text{E-}06$).

Visually, there was an excess of orthogroups with k close to 0, indicating a high number of orthogroups under pronounced relaxed selection (Figure 2.2). The number of orthogroups with $k < 1$ (1,387; relaxed selection prevailing) was slightly higher than the number of orthogroups with $k > 1$ (1,308; intensified selection prevailing), but not significantly so (Fisher's Exact Test, $P = 0.288$).

Of 2,702 orthogroups in the final dataset we conducted analyses of relaxed and intensified selection (in *RELAX*) and diversifying/positive selection (in *aBSREL* with branch-site model). We inferred 138 orthogroups to be under relaxed selection ($k < 1$, false discovery rate (FDR) < 0.10) in either *Coregonus* or *Salvelinus*, of which 122 were found in both *Coregonus* and *Salvelinus*. On the other hand, 105 orthogroups showed signals of intensified selection ($k > 1$, FDR < 0.10), of which 78 included both *Coregonus* and *Salvelinus*. The number of relaxed orthogroups in *Coregonus* and *Salvelinus* was significantly higher compared to the number of intensified orthogroups (one-sided Fisher's Exact Test, $P = 0.035$). Of the 2,702 orthogroups, branch-site selection analyses inferred 111 orthogroups as being under significant diversifying/positive selection (FDR < 0.10), of which 92 included both *Coregonus* and *Salvelinus*. Thus, these orthogroups harbour a proportion of sites with significantly elevated dN/dS ($= \omega$) values in at least one of the foreground branches leading to *Coregonus* or *Salvelinus* taxa.

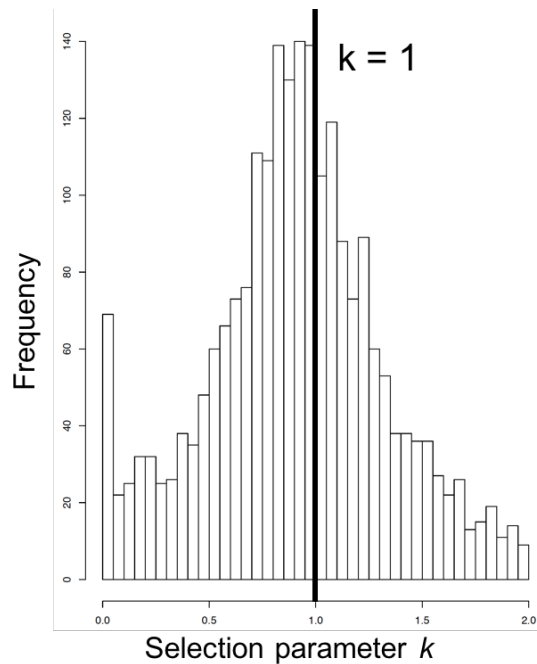


Figure 2.2 – Histogram of the distribution of selection parameter k values (exponent of dN/dS ratio, i.e. k in ω^k) from RELAX analysis in the 2,702 orthogroups. The values shown are for *Coregonus* and *Salvelinus* compared to the other five salmonid species and the outgroup (pike). The neutral expectation of $k = 1$ is shown as a vertical line. k values above 2 are omitted for visibility purposes in the above plot (416 orthogroups have k values above 2; the maximum possible k value is 50).

After averaging selection parameter values for each gene ontology (GO) term, 13 of 1,478 GO terms showed significant deviations from the null expectation of $k = 1$ (Wilcoxon signed-rank tests: $P < 0.05$; Figure 2.3). Eight of these had significantly elevated k values, indicating intensified selection. The other five had significantly lowered k values, which is evidence for relaxed selection. The GO term enrichment results agreed with the general shift of selection (distribution of k) in all orthogroups. The GO terms ‘carbohydrate metabolic process’ and ‘obsolete acyl-carrier-protein biosynthetic process’, for example, were also present in the orthogroups under intensified selection. The ‘ATPase activity’ and ‘proton transmembrane transport’ GO terms were also found among orthogroups under relaxed selection. Other deviating GO terms were ‘DNA repair’ and ‘protein deubiquitination’, with evidence for intensified selection, and ‘exocytosis’ and ‘protein dephosphorylation’, with evidence for relaxed selection (Figure 2.3).

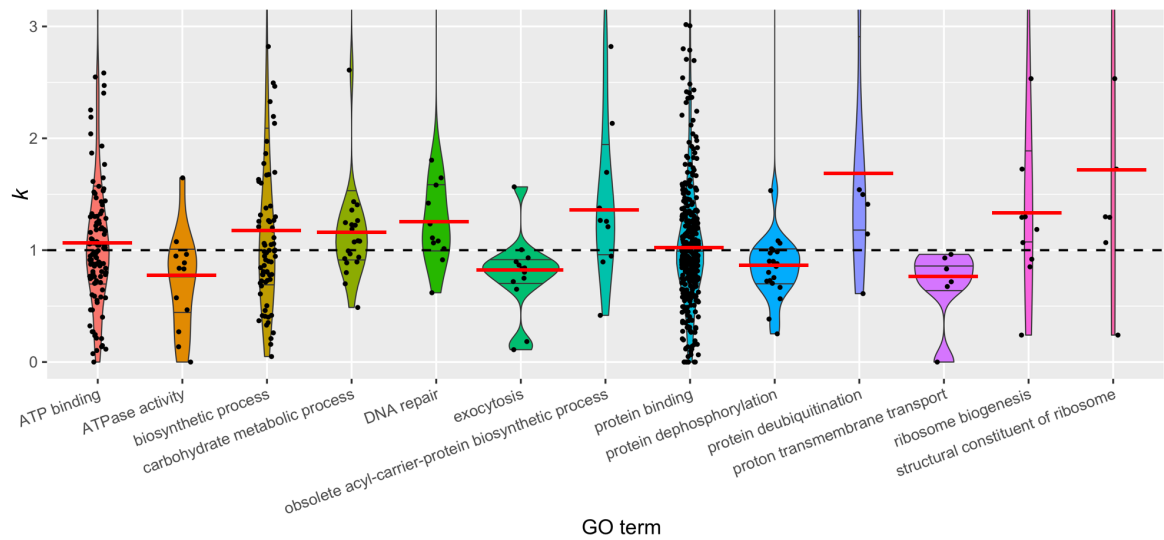


Figure 2.3 – Selection parameter k distributions for the gene ontology (GO) terms that deviated from the null expectation in the total set of 2,702 orthogroups. k values are selection parameter values (exponent of dN/dS ratio, i.e. k in ω^k) in *Coregonus* and *Salvelinus* compared to the five other salmonid species and the outgroup (pike). The red horizontal bars indicate the medians of k .

2.4.2 Gene functions under relaxed selection

Blast2GO annotation and UniProt/Swiss-Prot literature research on the orthogroups under relaxed selection identified gene functions with potential relevance for the diversification process in charr and whitefish. Such functions included visual perception (e.g. ‘peripherin-2-like’), gene and gene product regulation (e.g. ‘E3 ubiquitin-protein ligase RNF128-like’), lipid metabolism (e.g. ‘calcium-independent phospholipase A2-gamma-like’), muscle and heart growth (e.g. ‘dual specificity protein phosphatase 6’), locomotion (e.g. ‘serine/threonine-protein phosphatase PP1-beta catalytic subunit’), and immunity (e.g. ‘adaptor-related protein complex 1’, ‘natterin-3-like’), but also genes with a role for various nervous system processes (e.g. ‘POU domain, class 4, transcription factor 3-like’).

We observed compelling trends of GO term enrichment (one-tailed Fisher's Exact Tests, uncorrected $P < 0.05$ but FDR > 0.10) in the orthogroups under relaxed, intensified, and diversifying selection that largely agree with the research literature on the genes contained in those orthogroups (Figure 2.4, Table S2.1). We found the 122 orthogroups under significant relaxed selection in *Coregonus* and *Salvelinus* were enriched for a total of 11 GO terms associated with transcriptional regulation, serine family amino acid metabolism, lipid metabolism, and oxidoreductase activity, amongst others (Table S2.1). The REVIGO redundancy analysis results showed transcriptional regulation, serine family amino acid metabolism, lipid metabolism, and acrosome reaction to be among the few non-

redundant GO terms (frequency and significance plot of non-redundant GO terms: Figure 2.4a, includes clustering by semantic similarity). Transcriptional regulation and serine family amino acid metabolism were the most frequent non-redundant GO terms. In total, six of 11 GO terms were found to be non-redundant.

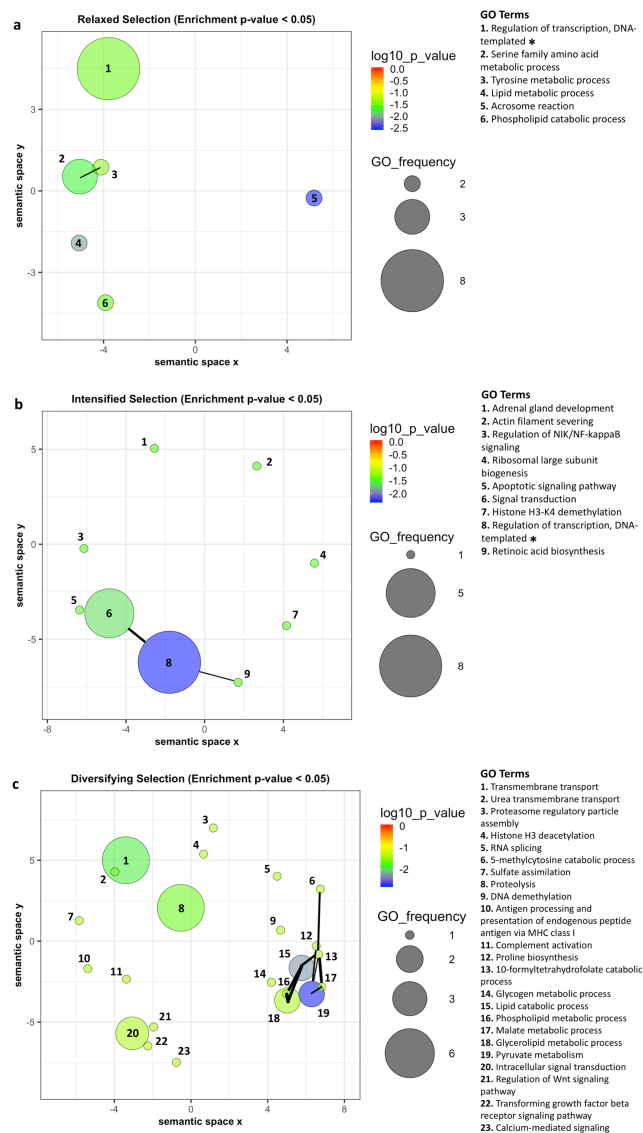


Figure 2.4 – Gene Ontology (GO) terms with trends of overrepresentation ($P < 0.05$) in a) the orthogroups under relaxed selection ($FDR < 0.10$), b) the orthogroups under intensified selection ($FDR < 0.10$, and c) the orthogroups under diversifying selection ($FDR < 0.10$). The orthogroups used are present in both *Coregonus* and *Salvelinus*. Bubble colour in the indicated colour spectrum corresponds to the $\log_{10} P$ value of overrepresentation (blue = higher significance, red = lower significance). Bubble size corresponds to the frequency of a GO term in the orthogroups under selection. Highly similar GO terms are linked by edges in the graph, where the line width indicates the degree of similarity. The bubble placement corresponds to the position along two semantic space axes based on SimRel clustering as described in Materials & methods (Supek et al. 2011). Asterisks indicate overlap with enriched GO terms in orthogroups under other types of selection.

Among the top ten enriched functions in relaxed orthogroups in both *Coregonus* and *Salvelinus*, behaviour and many neural function GO terms and KEGG pathways were found in gene set enrichment analyses (Figure 2.5). This is in agreement with the neural process orthogroups and serine family amino acid metabolism GO terms obtained in the GO term enrichment analysis above. The behaviour gene set was the only gene set that was significantly enriched after FDR correction (Figure 2.5). Other overrepresented functions included, for example, negative regulation of signalling, urogenital system development, the peroxisome pathway, vascular smooth muscle contraction, and the AGE-RAGE signalling pathway, which plays a major role in inflammation and infection.

2.4.3 Gene functions under intensified selection

The gene functions of the 78 orthogroups under intensified selection in both *Coregonus* and *Salvelinus* were found from literature search to be frequently involved in functions relevant for lipid and carbohydrate metabolism (e.g. ‘acetyl-CoA carboxylase beta’ and ‘endoplasmic reticulum mannosyl-oligosaccharide 1,2-alpha-mannosidase-like’, respectively) as well as neurological and bone development (e.g. neurological development: ‘synapsin-3’, bone development: ‘cathepsin K precursor’ and ‘paired like homeodomain 1’).

The orthogroups under intensified selection in *Coregonus* and *Salvelinus* were enriched for transcriptional regulation GO terms, but also for those associated with ubiquitine-related processes and steroid hormone receptor activity, amongst others. A total of 18 GO terms were overrepresented (Table S2.1). Transcriptional regulation and several signalling processes were the only high-frequency GO terms among the few non-redundant GO terms in the *REVIGO* analysis (frequency and significance plot of non-redundant GO terms: Figure 2.4b, includes clustering by semantic similarity). In total, nine of 18 GO terms remained after the *REVIGO* redundancy analysis.

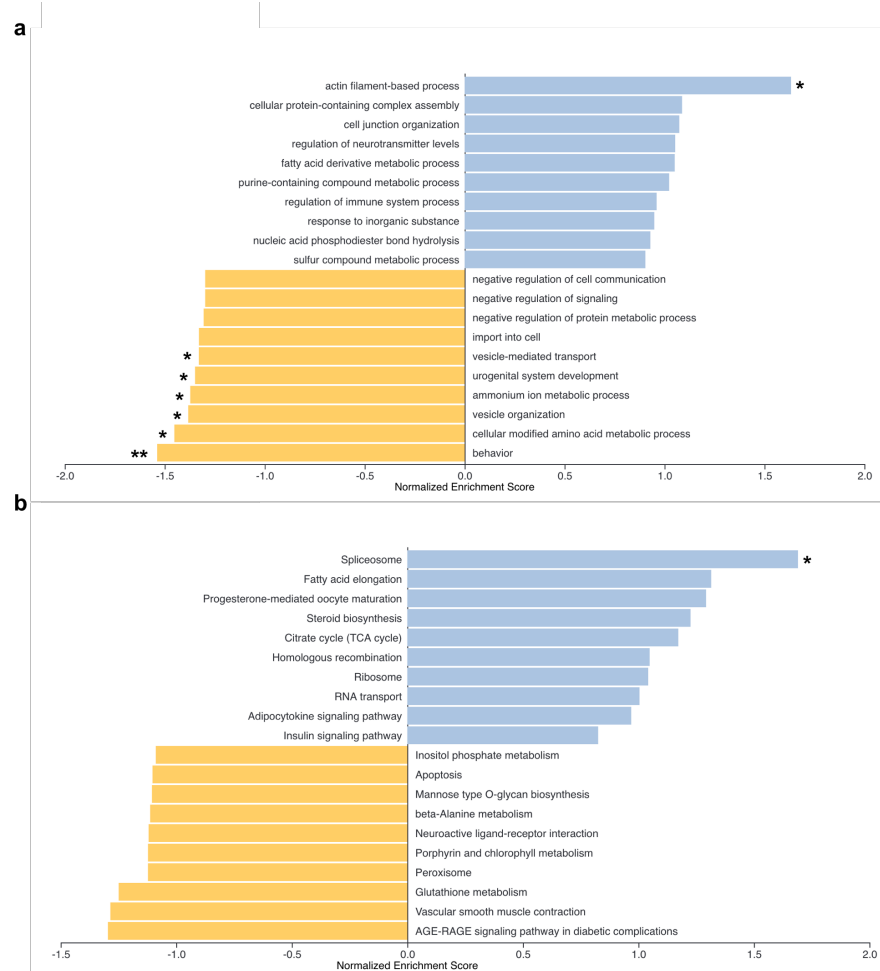


Figure 2.5 – Top ten enriched a) biological process GO terms and b) KEGG pathways based on a gene set enrichment analysis (GSEA) of all orthogroups in *Coregonus* and *Salvelinus* relative to non-diversifying taxa. Terms with negative scores (yellow) were inferred to be under relaxed selection and with positive scores (blue) under intensified selection. The normalised enrichment score is calculated from transformed k values of each gene in a gene set. ** indicates significant gene set enrichment after multiple correction ($FDR < 0.2$); * indicates significant gene set enrichment without multiple correction ($P < 0.05$).

In the gene set enrichment analysis of all intensified selection orthogroups present in *Coregonus* and *Salvelinus*, the ‘actin filament-based process’ GO term, the ‘spliceosome’ and several signalling KEGG pathways were among the top enriched functions (Figure 2.5). Other functions included ‘cellular protein-containing complex assembly’, ‘fatty acid elongation’, ‘progesterone-mediated oocyte maturation’, and ‘steroid biosynthesis’. Overall, the gene enrichment results (Figure 2.5) mostly agree with the GO term overrepresentations (Figure 2.4b).

2.4.4 Gene functions under diversifying selection

A large number of the 92 orthogroups under diversifying selection in *Coregonus* and *Salvelinus* were found in literature search to contain genes involved in regulation of gene expression, signal transduction and transmembrane transporter genes, but also immunity-related genes and a gene of the FOX set of genes, ‘FOX I1-ema’, a tissue-specific splicing factor important in otic placode formation and jaw development in zebrafish (Solomon et al. 2003).

Orthogroups under diversifying selection were enriched for GO terms associated with transmembrane transport, phospholipid metabolic processes, acetyl-CoA carboxylase activity, various lipid metabolic processes, regulation of Wnt signalling pathway, and RNA splicing, amongst others (Figure 2.4c, Table S2.2). A total of 47 GO terms were overrepresented, of which 23 remained after the *REVIGO* redundancy analysis. Pyruvate metabolism, several signal transduction processes, lipid metabolism, and transmembrane transport processes were shown to be amongst the non-redundant GO terms in the *REVIGO* analysis (frequency and significance plot of non-redundant GO terms: Figure 2.4c, includes clustering by semantic similarity). Compared to the orthogroups under relaxed or intensified selection (Figure 2.4a,b), the orthogroups under diversifying selection included a higher number of rather dissimilar low-frequency GO terms, apart from a cluster of similar metabolic GO terms (Figure 2.4c). Only one GO term, ‘DNA binding’, was underrepresented ($P < 0.05$); with zero occurrences in the orthogroups under diversifying selection in *Coregonus* and *Salvelinus* but 77 occurrences in all other orthogroups.

2.4.5 Overlap between selection types

We identified 9 orthogroups that showed both signals of relaxed selection (*RELAX*) and diversifying selection (*aBSREL*) and 12 orthogroups that showed both signals of intensified selection (*RELAX*) and diversifying selection (*aBSREL*) (Figure 2.6a, Table S2.2). The overlap between orthogroups under relaxed and diversifying selection was higher than expected by chance, but not significantly so (hypergeometric expectation: 3.6 vs observed 9; one-tailed Fisher's Exact Test: $P = 0.126$). The overlap between orthogroups under intensified and diversifying selection was significantly higher than expected by chance (hypergeometric expectation: 2.1 vs observed 12; one-tailed Fisher's Exact Test: $P = 0.004$).

Based on UniProt/Swiss-Prot gene information, the orthogroups with both signals of relaxed and diversifying selection are associated with functions such as immunity (5 of

9 orthogroups, e.g. ‘adaptor-related protein complex 1, mu 2 subunit (*ap1m2*)’), the nervous system (4 of 9 orthogroups, e.g. ‘protein kinase C epsilon type-like (*prkce*)’), muscle function (2 of 9 orthogroups, e.g. ‘solute carrier family 6 (neurotransmitter transporter), member 8 (*slc6a8*)’), blood pressure (1 of 9 orthogroups, ‘endoplasmic reticulum aminopeptidase 1-like (LOC106570844)’), and transcriptional regulation (1 of 9 orthogroups, ‘probable histone deacetylase 1-B (*hdac1-b*)’ (Table S2.2). This is in agreement with the more general GO term functions inferred using *Blast2GO* and associated tools (Figure 2.6b – biological process GO terms, Table 2.2), such as neurotransmitter transport, calcium-mediated signalling, antigen presentation, regulation of blood pressure, and serine family amino acid metabolism.

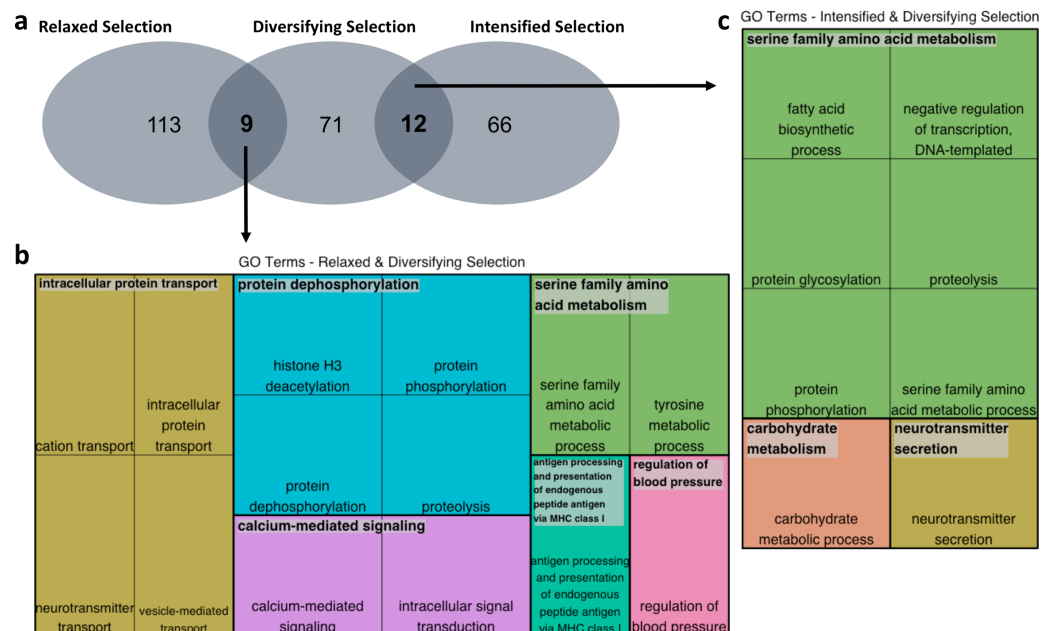


Figure 2.6 – Overlapping orthogroups under selection and biological process GO terms. a) Overlap between orthogroups under relaxed and diversifying or intensified and diversifying selection in both *Coregonus* and *Salvelinus*. b) Biological process GO terms of orthogroups under both relaxed and diversifying selection. c) Biological process GO terms of orthogroups under both intensified and diversifying selection. In b and c, rectangles are clustered into larger, coloured rectangles based on relationships among GO terms (Supek et al. 2011).

Based on UniProt/Swiss-Prot gene information, orthogroups with both signals of intensified and diversifying selection are associated with transcriptional regulation (4 of 12 orthogroups, e.g. ‘paired amphipathic helix protein Sin3a-like (*sin3a*)’), lipid metabolism (3 of 12 orthogroups, e.g. ‘acetyl-CoA carboxylase beta (*acacb*)’), nervous system function

(3 of 12 orthogroups, e.g. ‘synapsin-3 (*syn3*)’), carbohydrate metabolism (2 of 12 orthogroups, e.g. ‘alpha-2,8-sialyltransferase 8F-like (*st8sia6*)’), organelle function (2 of 12 orthogroups, e.g. ‘sterile alpha motif domain-containing protein 9-like (*samd9l*)’), bone growth (1 of 12 orthogroups, ‘cathepsin K precursor (*ctsk*)’), and immunity (1 of 12 orthogroups, ‘furin-1-like (*fur1*)’), amongst others (Table S2.2). Again, this agrees with the more general GO term functions inferred using *Blast2GO* and associated tools (Figure 6c – biological process GO terms, Table S2.2), such as neurotransmitter secretion, negative regulation of transcription, serine family amino acid metabolism, fatty acid biosynthesis, and carbohydrate metabolism.

2.5 Discussion

Our analyses of shared selection in the highly diversifying taxa *Coregonus* and *Salvelinus* identified genes and gene functions with deviating signatures of selection compared to five relatively less diversifying salmonid taxa and one non-salmonid species used as background. We identified more orthogroups under relaxed selection (k value < 1) than under intensified selection (k value > 1) (Figures 2.2 and 2.6). Further, we identified a set of 92 orthogroups that reflect signals of diversifying selection, 9 and 12 of which additionally experienced either relaxed or intensified selection, respectively (Figure 2.6, Table S2.2). We explored the underlying biological relevance and associated gene functions of these orthogroups under selection and observed trends of overrepresented GO terms in all three sets of orthogroups under selection (Figure 2.4, Table S2.1). Moreover, we identified 13 GO terms with k distributions deviating from the null expectation (Figure 2.3).

In our functional analyses of the above selected orthogroups, genes in orthogroups under relaxed selection were found to be involved in processes such as transcriptional regulation, nervous system function, muscle and heart growth, lipid metabolism, and immunity, while orthogroups under intensified selection were found to be more often involved in signalling processes, actin filament-based processes, RNA splicing, carbohydrate metabolism, transcriptional regulation as well as nervous system and bone development. Orthogroups under relaxed selection were enriched for ‘serine family amino acid metabolism’ while those under intensified selection tended to be enriched more strongly for various signalling pathways (Figure 2.4, Table S2.1). ‘Regulation of transcription’ was overrepresented in orthogroups under both relaxed and intensified selection. Orthogroups with signals of diversifying selection showed a clear trend towards enrichment for signal transduction, transmembrane transport, proteolysis, lipid metabolism,

and pyruvate metabolism GO terms (Figure 2.4, Table S2.1) and often contained genes involved in transcriptional regulation, development, lipid metabolism, and immunity. Therefore, our results imply different intensities of selection on different gene functions and processes.

2.5.1 Selection parameter distribution - shift to relaxation?

Our results agree with the repeated finding of a relaxation of selective constraint in rapidly diverging taxa (Yoder et al. 2010; Hunt et al. 2011; Calderoni et al. 2016; Škaloud et al. 2019). One possible cause behind this observation could be an increase of ecological opportunity due to the emergence of new habitat and a decrease in interspecies competition, as for example in caves (Calderoni et al. 2016) or within postglacial lakes (Siwertsson et al. 2010; Wellborn and Langerhans 2015; Schluter 2016; Recknagel et al. 2017; Doenz et al. 2018; Miller et al. 2019). The adaptation to life in lakes, before or in the course of (repeated) postglacial colonisations, such as in charr and whitefish, could be such an ecological opportunity, entailing a release of selective pressures on a substantial number of genes.

From the above results we can also conclude that relaxation of selective constraints does not affect the whole set of protein-coding genes to the same degree. Purifying selection and selective constraint still affect a large number of genes with functions that ensure an organism's integrity, survival, and reproduction (see orthogroups under intensified selection) (Shakhnovich and Koonin 2006). However, in charr and whitefish we observed a larger number of orthogroups under relaxed selection that therefore escaped the effects of purifying selection.

The observation of largely different GO terms in orthogroups under relaxed selection or intensified selection agrees with a large number of studies that demonstrated different selective pressures on different gene functions (Mitterboeck et al. 2017; Roux et al. 2017). Various functional and structural constraints govern the substitution patterns within genes. For example, amino acid alterations in some groups of proteins can easily make them dysfunctional (Roux et al. 2017), while other kinds of proteins are under less constraint to vary (Roux et al. 2014) or are even beneficial to fitness when being different from the majority of the population (i.e. negative frequency-dependent or balancing selection) (Du et al. 2011; Gagnaire et al. 2012; Yan et al. 2017). The observation of intensified selection on DNA repair in our dataset and the overrepresented GO terms related to oxidative stress and apoptotic signalling, for example, could highlight the importance of an effective response to environmental stress in rapidly diversifying taxa.

Also, carbohydrate and ribosome-related processes seem to be under strong selective constraint. On the other hand, relaxation of selective constraint on ATPase activity and proton transmembrane transport as well as various lipid metabolism genes, amongst others, could have led to the physiological variety observed in extant highly diversifying salmonids.

2.5.2 Gene functions under relaxed selection in *Coregonus* and *Salvelinus*

Many of the orthogroups under parallel relaxed selection, i.e. those that experienced a release of selective constraint in *Coregonus* and *Salvelinus*, have been linked to the divergence of these species along the benthic-limnetic axis within lakes found in earlier research (Bernatchez et al. 2016; Carruthers et al. 2018; Doenz et al. 2018; Varadharajan et al. 2018; Jacobs et al. 2019, 2020). Examples include the ‘peripherin-2-like’ gene, which has been implicated in visual adaptation in *S. namaycush* in Lake Superior (Perreault-Payette et al. 2017) and the ‘natterin-3-like’ gene, which plays a putative role as an immunopeptide and experienced pronounced expression divergence in Icelandic *S. alpinus* ecomorphs (Steinhäuser 2013; Gudbrandsson et al. 2015). Based on our analyses, we propose these are linked to the evolution of novel phenotypes and the variation in feeding and swimming behaviour, energy storage, and the release of competition in postglacial lakes (behaviour, muscle and heart growth, locomotion, lipid metabolism, nervous system function) as well as a release or change of parasite and pathogen burden (immunity genes) in postglacial lakes (Poulin 2016). This implies changes in selective pressure on these gene functions as a consequence of shifts in the ecology of highly diversifying salmonids.

2.5.3 Gene functions under intensified selection in *Coregonus* and *Salvelinus*

Genes under intensified selection are expected to have a strong role for survival and reproduction (i.e. fitness) of organisms (Wertheim et al. 2015). We found the orthogroups under intensified selection showed a high number of processes associated with key metabolic pathways (mainly of carbohydrates and lipids), signal transduction, and regulation of gene expression and the immune system, indicating strong selective pressures on these processes and functions in *Coregonus* and *Salvelinus*. The occurrence of immunity genes under intensified selection in our results would rather imply changing parasite and pathogen regimes rather than a complete release of pathogen burden (Xing and Lee 2006). Interestingly, our results also indicated strong selective pressures on actin filament-based processes and RNA splicing as well as protein modification, which are functions with a crucial role in development (Xing and Lee 2006; Rottner and Stradal

2011; Reimand et al. 2015; Tan et al. 2015). Among these genes, some were found to be directly linked to differences in bone development (e.g. ‘paired like homeodomain 1’ and ‘cathepsin K precursor’) (Ahi et al. 2014; Ahi 2016; Beck et al. 2019). We speculate that these developmental process genes may be relevant in the extreme morphological diversity that can be found among closely related *Coregonus* species and particularly *Salvelinus* species (Ahi et al. 2014; Ahi 2016; Beck et al. 2019; Jacobs et al. 2019; Rougeux et al. 2019; Striberny et al. 2019).

2.5.4 Gene functions under diversifying selection in *Coregonus* and *Salvelinus*

Diversifying (i.e. positive) selection on genes, which in this context means that selective pressure favours amino acid polymorphism in a particular gene, affected 92 orthogroups in *Coregonus* and *Salvelinus* relative to the non-diversifying taxa. The enrichment of lipid, glycerolipid, and pyruvate metabolism genes is a strong indication that these metabolic processes experienced diversification/positive selection in *Coregonus* and *Salvelinus*, which could imply a role in ecological diversification and habitat shifts across the depth axis in lakes. Lipid metabolism has been commonly found to be under diversifying selection in freshwater fishes, such as sticklebacks (Bruneaux et al. 2013) and salmonids (Varadharajan et al. 2018; Chen et al. 2019). Also, diversifying/positive selection on transmembrane transport, for example, has been shown repeatedly in teleost fish adapting to the freshwater environment, such as sculpin and alewife, putatively because of its role in osmoregulation (Dennenmoser et al. 2017; Velotta et al. 2017).

2.5.5 Overlap between selection types

A combination of approaches, such as *aBSREL* and *RELAX* as we apply here, has been shown to reliably infer genes under selection and is also able to appropriately distinguish between positive selection and relaxed selection, which are hard to disentangle using single conventional methods (Suzuki and Nei 2004; Nozawa et al. 2009; Jordan and Goldman 2011). This is particularly important since it has been known for some time that the number of false positives in tests for positive or diversifying selection can be substantial (Zhang 2004; Zhang et al. 2005; Yang 2007; Markova-Raina and Petrov 2011).

Overlapping sets of orthogroups under selection inferred with different analytical approaches, as we resolve here, can therefore implicate consistent, strong, and specific signals of selection. For example, orthogroups with signals of relaxed and diversifying selection likely experienced a release of purifying selection, with the potential for evolutionary diversification (Hunt et al. 2011; Wertheim et al. 2015; Shinde et al. 2019).

One of the orthogroups with the most extreme signal of relaxed and diversifying selection was ‘adaptor-related protein complex 1, mu 2 subunit (*ap1m2*)’ (Table 2.2). Apart from a role in endothelial and intestinal immunity (Takahashi et al. 2011), *ap1m2* plays a central role in the development of endoderm-derived organs (e.g. stomach and intestine) in zebrafish (Zizioli et al. 2010). Given the differing parasite pressure (Siwertsson et al. 2016; Brunner et al. 2017) and the diversity of trophic niches in lake systems (Lu and Bernatchez 1999; Hooker et al. 2016), genes under reduced selective constraint could play a crucial role in phenotypic diversity in the often lake-dwelling *Salvelinus* and *Coregonus*. Another gene showing a strong signal of relaxed and diversifying selection was ‘solute carrier family 6 (neurotransmitter transporter), member 8 (*slc6a8*)’, which regulates creatine uptake in tissues with high energy demands such as muscle and brain tissue (Braissant et al. 2007; Borchel et al. 2014; Hooker et al. 2016). Recent research on *O. mykiss* and *C. maraena* found two copies of this gene, which were either mainly expressed in muscle or kidney tissue, with the strongest expression overall in muscle (Borchel et al. 2019). Changes of selection regimes on these functions are expected, given that *Salvelinus* and *Coregonus* experienced a diversification in trophic and swimming behaviour, ecology, and life history (e.g. *Salvelinus* and *Coregonus* can be free swimming, lake bottom dwelling, or restricted to the littoral zone of lakes). Our analyses independently corroborate an important functional role of this gene in highly diversifying salmonids.

Orthogroups showing both intensified and diversifying selection likely constitute targets of "true" positive selection rather than relaxed selection misinterpreted as positive selection. Genes in these orthogroups are thus under stronger selective pressure and are potentially of higher importance to survival and overall fitness (Wertheim et al. 2015) in our analyses of *Coregonus* and *Salvelinus*. Two of the most strongly significant orthogroups under intensified and diversifying selection involved the genes ‘synapsin-3 (*syn3*)’ and ‘cathepsin K precursor (*ctsk*)’. *syn3* is involved in neurotransmitter secretion (Feng et al. 2002) and plays a crucial role in early neuronal differentiation and in neuronal progenitor cell development in fish (Garbarino et al. 2014) as well as in mammals (Kao et al. 2008). *ctsk*, encoding a collagenase, has an important function in bone modelling and remodelling as well as bone homeostasis in vertebrates (Saftig et al. 1998; To et al. 2015) and is differentially expressed in *S. alpinus* (Beck et al. 2019). Both *syn3* and *ctsk* are among the genes experiencing the highest selective pressure in *Coregonus* and *Salvelinus* as compared to other salmonid taxa, which is an indication of their importance for survival in these two genera and warrant further research.

Not only specific genes but also the gene functions associated with orthogroups under selection were found to have biological relevance. Changes in heart and muscle function, immunity genes, lipid metabolism, and transcriptional regulation have previously been implicated in diversification in salmonids (Kapralova et al. 2013; Conejeros et al. 2014; Johnston et al. 2014; Jacobs et al. 2017; Roux et al. 2017; Gillard et al. 2018; Larson et al. 2019), cichlids (López-Fernández et al. 2013), ants (Roux et al. 2014), and tonguefishes (Du et al. 2011). Therefore, our findings from molecular evolution agree with evidence from ecomorphological divergence of *Coregonus* and *Salvelinus* of immune response, feeding, metabolism, and locomotion (Bernatchez et al. 2010; Jeukens et al. 2010).

2.5.6 Whole-genome duplication and potential effects on relaxed selection

Multiple lines of evidence suggest that whole-genome duplication (WGD) events give rise to vast amounts of genetic material that can subsequently experience elevated substitution rates and relaxation of selective constraints (Lynch and Coneri 2000; Kondrashov et al. 2002; Brunet et al. 2006). The lineage ancestral to salmonids experienced a WGD around 80 – 103 Mya (Macqueen and Johnston 2014; Lien et al. 2016; Robertson et al. 2017). It has been speculated that this may have contributed to new phenotypic innovations and to the ecological success of salmonids (Taylor et al. 2001; Hoegg et al. 2004; Phillips et al. 2009; Crête-Lafrenière et al. 2012; Macqueen and Johnston 2014), although a contribution to diversification rates is contentious (Santini et al. 2009; Macqueen and Johnston 2014). While conventional approaches for the detection of positive and relaxed selection based on orthologues in salmonids would potentially be biased in the case of duplicated and non-rediploidised genes, approaches based on orthogroups, as we applied here, alleviate this issue by explicitly comparing selection shared in *Coregonus* and *Salvelinus* relative to all other salmonid lineages (Murrell et al. 2015). This approach circumvents the difficulty of inferring orthologues in taxa that experienced multiple duplications and unequal rediploidisation.

While genes duplicated during the salmonid WGD might be more susceptible to relaxed selection (Innan and Kondrashov 2010; Wertheim et al. 2015), charr and whitefish are in different subfamilies and not sister genera (Figure 1) (Macqueen et al. 2017; Lecaudey et al. 2018). Therefore, based on evolutionary relationships, they should have rediploidisation patterns more similar to their sister genera than to each other and there is no reason to suspect unequal rediploidisation among genera to be the cause of the relaxed selection signals we identified in *Coregonus* and *Salvelinus*. It is more likely that these two

genera experienced similar selective pressures due to common diversification patterns, environmental conditions, or other common ecological or life history traits. Our results from molecular evolution analyses reflect these shared ecological and evolutionary patterns. However, recent results on gene expression in duplicated genes of salmonids have shown that relaxed selection often occurs in downregulated gene duplicates (Sandve et al. 2018), which might be associated with dosage balance effects (Rosado and Raikhel 2010; Birchler and Veitia 2012). Therefore, future research should ideally analyse selection in concert with gene expression and duplication.

2.6 Conclusions

Our findings suggest that there is some evidence for a parallel relaxation of selective constraint in the repeatedly diversifying salmonid lineages *Coregonus* and *Salvelinus* compared to all other salmonid lineages, which are known to be less diversifying. Genes in orthogroups under relaxed selection are involved in functions with a potential role in the rapid diversification and ecological adaptation that can be observed in wild populations of *Coregonus* and *Salvelinus* and experienced a release of selective pressure, including genes involved in behaviour, muscle function, and infection. On the other hand, orthogroups under intensified selection, such as various signal transduction and regulatory genes, are under stronger selective pressure and are consequently expected to have higher fitness effects in charr and whitefish compared to other salmonids. An independent analysis of orthogroups under diversifying selection showed an enrichment for signal transduction genes and GO terms of various metabolic processes, while genes involved in DNA binding were underrepresented. Orthogroups with both signals of relaxed and diversifying or intensified and diversifying selection can give further hints as to what selective processes govern the evolution of these genes and are also important candidates for gene sets under particularly weak (relaxed) selective pressure (while still experiencing molecular diversification), or strong (intensified) selective pressure. The discovered orthogroups under selection are a valuable resource for further investigations into the importance of certain genes for rapid diversification in salmonids and beyond.

Chapter 3: Pitfalls and virtues of population genetic summary statistics: detecting selective sweeps in recent divergences

3.1 Abstract

During evolution, genomes are shaped by a plethora of forces that can leave characteristic signatures. A common goal when studying diverging populations is to detect the signatures of selective sweeps, which can be rather difficult in complex demographic scenarios, such as under secondary contact. Moreover, the detection of selective sweeps, especially in whole-genome data, often relies heavily on a narrow set of summary statistics that are affected by a multitude of factors, frequently leading to false positives and false negatives. Simulating genomic regions makes it possible to control these demographic and population genetic factors. We used simulations of large genomic regions to determine how different secondary contact and sympatric speciation scenarios affect the footprint of hard and soft selective sweeps in the presence of varying degrees of gene flow and recombination. We explored the ability of an array of population genetic summary statistics to detect the footprints of these selective sweeps under specific demographies. We focussed on metrics that do not require phased data or ancestral sequences and therefore have wide applicability. We found that a newly developed beta diversity measure, $\bar{B}^{GD}$, outperformed all other metrics in detecting selective sweeps and that F_{ST} also performed well. High accuracy was also found in $\Delta\pi$ and genotype distance-derived metrics. The performance of most metrics strongly depended on factors such as the presence of an allopatric phase, migration rates, recombination, population growth, and whether the sweep was hard or soft. We provide suggestions for locating and analysing the response to selective sweeps in whole-genome data.

3.2 Introduction

Evolutionary processes commonly act on genomes in a concerted manner and often leave characteristic signatures. Frequently studied examples include positive selection in the form of soft and hard selective sweeps (Hermisson and Pennings 2005; Pritchard et al. 2010), background selection (Charlesworth et al. 1995), recombination (Bamshad and Wooding 2003), and introgression (Hufford et al. 2013; Moest et al. 2020). While these responses and processes leave typical signatures on patterns of diversity across the genome, the signatures can become eroded and are therefore increasingly difficult to

capture over time. Many studies have attempted to screen for signatures (such as those of positive selection) in ongoing or recent evolutionary events in nature (Gagnaire et al. 2013; Malinsky et al. 2015; Torres-Dowdall et al. 2015; Lamichhaney et al. 2018; Rettelbach et al. 2019; Jacobs et al. 2020) and in experimental evolution settings (Dunham et al. 2002), such as in the form of evolve-and-resequence experiments (E&R) (Long et al. 2015; Schlötterer et al. 2015).

Because evolutionary responses are variable, complex, and ultimately unknowable in natural settings, simulating genomic regions *in silico* is an approach to dissecting the influence of specific evolutionary forces on genomes. This makes it possible to control evolutionary forces rigorously and determine the effect of particular modifications. Simulation studies have led to important insights about the genomic footprints of evolutionary forces such as natural selection (Lohmueller 2014; Matthey-Doret and Whitlock 2019). Empirical studies on population genomics and evolutionary biology can use simulations to obtain informed estimates of genome signatures under given evolutionary scenarios and compare the simulated and empirical patterns (Enard et al. 2014; Rettelbach et al. 2019).

Despite the incorporation of various demographic models in *in silico* and empirical studies, many demographic scenarios have so far received comparatively little attention. An example is secondary contact, in which two populations become sympatric again after an allopatric phase. This scenario has been shown to be particularly common in recent divergences, many of which are affected by the expansion and retreat of glaciers, including many temperate region animal model systems in ecological and evolutionary genomics (Schluter 1996; Pettengill and Moeller 2012; Gagnaire et al. 2013; Elmer 2016). Other examples include secondary contact after range expansion in ecotypes of the plant model systems *Mimulus* sp. and *Clarkia* sp. (Streisfeld and Kohn 2005; Pettengill and Moeller 2012; Kenney and Sweigart 2016). Secondary contact scenarios can be rather complex because they include elements of allopatric and sympatric speciation in varying proportions (Jacobs et al. 2020), complicated by processes such as reinforcement (Noor 1999) and character displacement (Brown and Wilson 1956), which accentuate differences accumulated in allopatry. Sympatric speciation, on the other hand, does not include an allopatric phase. Instead, divergence takes place under actual or potential gene flow between incipient species or subpopulations (Foote 2018).

While whole-genome sequence data are the most informative for discerning evolutionary processes, screening such data sets for genomic responses to selection and recent divergences can be challenging. Very large amounts of data are difficult to analyse with model-based approaches, such as likelihood-based and Bayesian sweep-finding

methods (Nielsen et al. 2005; Jensen et al. 2008; Pavlidis et al. 2013; Huber et al. 2016), due both to computational limitations and to difficulties in comparing across study systems (Hämälä and Savolainen 2019; He et al. 2019). To reduce sequencing costs and maximise numbers of individuals in ecological and evolutionary model systems, in many cases low coverage sequence data are used to infer genotype likelihoods, in which all possible genotypes at each nucleotide position in the genome are assigned a likelihood value. Currently most model-based methods are not suitable for genotype likelihoods and rely on individual-level called genetic variants, where polymorphisms in individuals are known with high certainty (Nielsen et al. 2011; Korneliussen et al. 2014). Additionally, sequence data of low to moderate coverage are usually not suited for phasing or haplotype inference, which are prerequisites for many metrics to infer selection or other evolutionary processes (Sabeti et al. 2007; Alachiotis et al. 2012; Ferrer-Admetlla et al. 2014). Lastly, many commonly used metrics to infer selection require a polarised site frequency spectrum, which requires knowledge about the ancestral sequence (Fay and Wu 2000) that may or may not be available.

To overcome these limitations, summary statistics are commonly used, often in concert, to screen for characteristic signals of selective sweeps. Population genetic metrics applied across the genome are widely used by evolutionary biologists to determine regions under selection compared to background divergence, such as *Tajima's D* (Tajima 1989), which can be used to detect positive selection, and F_{ST} (Wright 1949), which tests for differentiation between populations but can also be used to screen for differential selection (Beaumont 2005; Holsinger and Weir 2009). The candidate genomic regions can then be evaluated for functional genes and potential function-altering substitutions (Ng and Henikoff 2006; Barrett and Hoekstra 2011) or combined with evidence from phenotypic data (Stinchcombe and Hoekstra 2008; Barrett et al. 2019) to further infer their evolutionary significance.

A major drawback of applying population genetic summary statistics to regions of the genome is that they can be influenced by processes other than natural selection, such as gene flow, recombination, population structure, and demography (Stajich and Hahn 2005). In the case of F_{ST} for example, several studies have demonstrated how false positives can result from the demographic history of a population (Lotterhos and Whitlock 2014), differential introgression (Gosset and Bierne 2013), hierarchical population structure (Excoffier et al. 2009), variation in recombination rates within the genome (Lotterhos 2019), and background selection (Matthey-Doret and Whitlock 2019). Similarly, *Tajima's D* has been shown to produce biased estimates of selection depending on

population history (Stajich and Hahn 2005), population structure (Moeller et al. 2007), and recombination rate variation (Thornton 2005). The influence of demography, gene flow, and genomic structure on other summary statistics, particularly those that have been developed recently, is not well understood (Ravinet et al. 2017).

Metrics that draw upon whole-genome data commonly make use of the following genomic properties: 1) The site frequency spectrum, which is the distribution of allele frequencies at a number of loci in the genome, shifts in response to selection due to genomic hitchhiking (Nielsen et al. 2005) but also can be influenced by demographic factors (Stajich and Hahn 2005). 2) The distribution of genotype distances g_{kl} is a recent approach that essentially captures sequence divergence in particular genomic regions and has to date only been assessed using a small set of evolutionary scenarios (Kern and Schrider 2018). 3) Multilocus genotype frequency metrics, which are similar to haplotype frequency statistics for unphased data (Messer and Petrov 2013; Garud et al. 2015), can also be used to infer sweeps but have been shown to have little power in the case of soft sweeps (Harris et al. 2018; Kern and Schrider 2018). 4) Linkage disequilibrium can be used to detect sweeps, but metrics have primarily relied on phased data (Kelly 1997; Kim and Nielsen 2004), while recently these have been extended to unphased data (Schlamp et al. 2016; Kern and Schrider 2018). However, Kim and Nielsen (2004) pointed out that the relative power to detect sweeps is weaker using linkage disequilibrium compared to methods based on allele frequencies (e.g. the site frequency spectrum). 5) Differentiation between populations can also be used to infer differential selection but can be confounded by demographic processes (Chen et al. 2010; Lotterhos and Whitlock 2014). Until now, a full assessment of the relative performances of all of these metrics to detect selective sweeps under a range of evolutionary scenarios has not been made.

To better understand the characteristic patterns of selective sweeps and the relative performance of summary statistics under different demographic scenarios, we performed simulations of how genomes diverge between populations under an array of evolutionary and demographic scenarios. These include positive selection in one population and negative selection of varying degrees in the other population, reflecting differential selection and adaptation as a consequence of ecological specialisation (Poisot et al. 2011). We did this in the versatile simulation environment SLiM (Messer 2013; Haller and Messer 2017), which allows for bespoke adjustment of genetic and evolutionary processes. Sixteen summary statistics were assessed with regard to their ability to detect the selective sweeps that we simulated: π , $\Delta\pi$, $\hat{\theta}_W$, *Tajima's D*, the variance, skewness, and kurtosis of the genotype distance distribution g_{kl} (Kern and Schrider 2018), the multilocus genotype

frequency-based $nDiplos$, J_1 , J_{12} , and J_2/J_1 (Messer and Petrov 2013; Garud et al. 2015; Kern and Schrider 2018), the genotypic linkage disequilibrium-based $diploZ_{NS}$ and $diplo\omega$ (Kelly 1997; Kim and Nielsen 2004; Kern and Schrider 2018) as well as $H - scan$ (Schlump et al. 2016), F_{ST} , and a recently developed beta diversity measure based on information theory (Reeve et al. 2016) newly adapted to this genomic context, $\bar{B}^{GD}$ (White 2019).

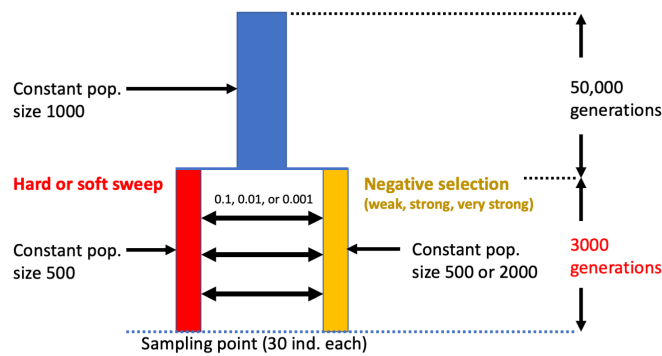
We compared the relative accuracy of the 16 metrics above in predicting the location of the simulated selective sweep under the series of different evolutionary and demographic histories. We calculated accuracies based on 1) the correct identification of the location at which the sweep occurred, detected as peaks or valleys of summary statistics, and 2) significant signals of selection when comparing to distributions of summary statistics derived from neutral simulations. We found large differences in performance among metrics, with some summary statistics outright failing to detect the location of selection under particular demographic scenarios. The newly developed genetic beta diversity measure, $\bar{B}^{GD}$, outperformed all other metrics we assessed, including F_{ST} . Based on our simulation results, we provide some guidance as to which metrics are useful for sweep prediction, especially in the case of populations with complex, recent histories of divergence.

3.3 Materials & methods

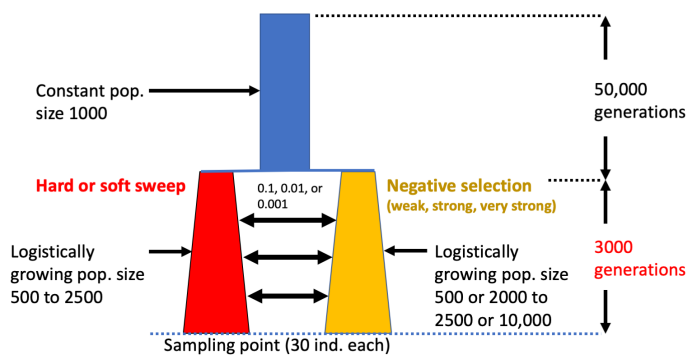
3.3.1 Simulations

The scripts for all simulations were written in the Eidos language (Haller 2016) for simulation in SLiM 2 (Haller and Messer 2017). A total of 144 scenarios were simulated: 96 scenarios that included a selective sweep and 48 scenarios without selection (general outline: Figure 3.1; all parameter combinations: Table S3.1). These scenarios modelled either sympatric speciation (i.e. speciation-with-gene-flow and no isolation in the initial stages of divergence) or secondary contact (i.e. an allopatric phase during the initial stage of population divergence followed by gene flow at secondary contact). The parameter settings were meant to reflect recent divergence scenarios, for example speciation following colonisations of lakes or other disjunct habitats. Generations in all simulations were modelled as being discrete and the simulation duration was 53,000 generations. Individuals in post-divergence populations are offspring of pre-divergence populations.

a) Sympatric Divergence (Constant Population Size)



b) Sympatric Divergence (Growing Population Size)



c) Secondary Contact (Constant Population Size)

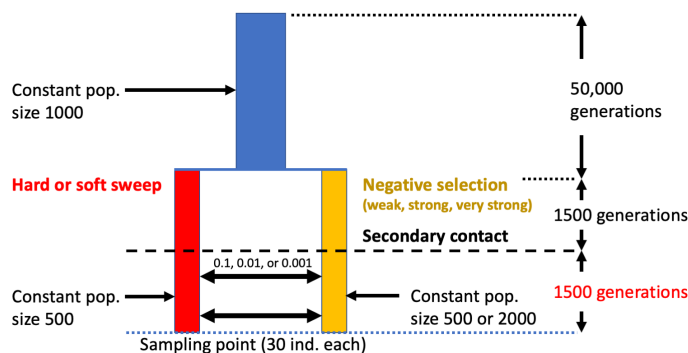


Figure 3.1 – General outline of simulated scenarios. a) Sympatric divergence scenario where gene flow occurs directly after the population split and continues until the samples are taken. In this case, population sizes remain constant after divergence. b) Sympatric divergence scenario with logistic population growth. c) Secondary contact scenario where there is no gene flow directly after the population split. Gene flow continues after secondary contact of populations 2 and 3 until the samples are taken. All combinations of parameters were included in the simulations for a total number of 144 scenarios, including 96 selective sweep scenarios and 48 neutral scenarios.

To build up genetic variation, all scenarios involved a 50,000-generation pre-divergence period for a population of 1000 individuals (hereafter 'population 1'). Sympatric speciation simulations (hereafter 'scenario 1') then involved divergence into two populations initially of 500 individuals, with each population experiencing low levels of gene flow for another 3000 generations until the simulations ended. In some contexts, sympatry with low levels of gene flow rather than panmixia is often called parapatry (Gavrilets 2004), but we will call it sympatry in the following since ecological divergence of geographically sympatric populations typically entails a reduction in gene flow of varying rates (Rundle and Nosil 2005; Nosil 2012), while parapatry could also imply overlapping but not completely sympatric populations in a biogeographic sense (Schuler et al. 2006). Secondary contact experiments involved the same divergence into two populations of 500 individuals; thereafter each population experienced a 1,500-generation allopatric phase, followed by a 1,500-generation period with gene flow until the simulations ended.

Selection: In all scenarios that included selection (the neutral simulations are described below), the pre-divergence population ('population 1') was not subject to any selection, one of the post-divergence populations (hereafter 'population 2') experienced relatively strong positive selection (selection coefficient $s = 0.1$; relative fitness of offspring per generation $f = 1 + s = 1.1$ for both homozygote and heterozygote carriers of the selected mutation; see http://benhaller.com/slim/SLiM_Manual.pdf) at two loci (one in a region with low and one in a region with standard recombination rates), while the other population (hereafter 'population 3') experienced negative selection of varying strength caused by the same mutations that are positively selected in population 2. Depending on the relative fitness, f , of mutations in population 3, secondary contact scenarios are designated as 'scenario 2' ($f = 1 / \text{relative fitness of population 2} = 0.909$), 'scenario 3' ($f = (1 / \text{relative fitness of population 2}) - 0.1 = 0.809$), or 'scenario 4' ($f = (1 / \text{relative fitness of population 2}) - 0.5 = 0.409$, i.e. substantially reduced fitness). The chosen selection coefficient of $s = 0.1$ has been shown to be realistic and common in a meta-analysis of selection coefficients in natural populations (Thurman and Barrett 2016). The rationale for including stronger negative selection in scenarios 3 and 4 was to mimic the effects of an allele of one population being confronted with a different genomic background in another population upon secondary contact, which can lead to lower fitness in hybrids or even hybrid sterility/inviability (Fang et al. 2012; Crespi and Nosil 2013; Sweigart et al. 2019).

Demography: In all scenarios the migration rates, m , the number of exchanged individuals per generation, were set to 0.1, 0.01, or 0.001 across different simulations. Migration rates here can be interpreted as the fraction of individuals of one population

contributing to offspring in the next generation of the other population. All scenarios were modelled either with population sizes held constant at 500 individuals per population after population divergence or population sizes logistically growing from 500 individuals to 2500 individuals per population. Population sizes in population 3 were either the same as population 2 or higher, with 2000 individuals in the case of constant population sizes or growing to 10,000 individuals in the case of increasing population sizes. All combinations of parameter values and scenarios were tested. The effective population sizes and parameter settings used are meant as a compromise between more extreme sizes and values and account for limitations due to computational feasibility.

Selective sweep: Scenarios were modelled with 'population 2' experiencing a hard sweep, a soft sweep, or no selection (neutral evolution). Soft sweeps were modelled as selective sweeps from standing genetic variation in 'population 2', i.e. the mutation was introduced in 400 of 1000 individuals at generation 50,000 (immediately before divergence into populations 2 and 3) with a starting frequency of 0.4 in both post-divergence populations, which ensures that the selected allele is neither lost nor fixed, while at the same time making sure that starting points are equal across repetitions. Hard sweeps were modelled as one single point mutation in one individual in population 2. The onset of selection was always at generation 50,001 and continued to the end of the simulation. The selected mutations were modelled with a dominance coefficient, $h = 1.0$, denoting complete dominance, which is thought to be common in newly arising beneficial mutations (Haldane 1927; Turner 1977; Charlesworth 2017).

Genome structure: Genomes were simulated with a size of 4,000,050 bp. These consisted of two recombination regions of 2,000,025 bp each: one 'standard' recombination rate region (or probability) representative of many fish species, $r = 1\text{E-}7$, and one low recombination rate region with $r = 1\text{E-}8$ (Stapley et al. 2017). The recombination rates are the probabilities of recombination per nucleotide site per generation. The mutation rate in all simulations was $1\text{E-}8$ mutations per position per generation. The selective sweep mutations were always located in the centre of each recombination rate region, at positions 1,000,013 bp (in the standard recombination rate region) and 3,000,038 bp (in the low recombination rate region). The two recombination rate regions were completely independent (i.e. a selective sweep in one region did not affect the patterns in the other region) and can therefore be considered as separate chromosomes.

In total, simulations were replicated 100 times for each of the 144 different combinations of parameter values and scenarios (Table S3.1). The simulated genomes were output as files in both ms (Hudson 2002) and vcf format, depending on analysis software requirements and computational efficiency of the summary statistic analyses.

3.3.2 Summary statistics

We focussed on summary statistics that are computationally efficient, easy to apply, and do not require phased data or ancestral sequence information. Since machine learning methods to infer selective sweeps, such as evoNet (Sheehan and Song 2016) and S/HIC or diploS/HIC (Schridder and Kern 2016; Kern and Schridder 2018), usually require the simulation of training and testing data sets, drastically increasing the computational investment, we did not include these in the present study. Based on randomly sampling 30 individuals each of population 2 and population 3 from generation 53,000 of the simulations, we calculated summary statistics in 50 non-overlapping windows of 80,001 bp size using the ms or vcf simulation output files. The window size of 80,001 bp was found to lead to good resolution of signatures of selection after initial trials of various window sizes between 50,000 bp and 100,001 bp. Non-overlapping windows were used due to the selected allele being located in the centre of the central window of each recombination region, so problematic edge effects (i.e. mutation at the edge of a window leads to two minor peaks rather than one major peak) of selective signatures on empirical data can be excluded. Non-overlapping windows also allowed for faster computation times. To avoid biases due to definition of windows, we additionally defined windows using a total of 22 SNP-based windows (11 per recombination rate region) holding 21 SNPs per window, which were held constant and are based on the minimum number of SNPs present in each of the simulation runs. The chosen number also ensured that the central window had the selected mutation in its centre and is therefore compatible with the 1-in-1 accuracy method described below. These windows holding 21 SNPs each were calculated based on individual SNP metric values for each position using a custom R script. Individual SNP metric values could not be calculated for the linkage disequilibrium-based metrics $diploZ_{NS}$ and $diplo\omega$ of the diploSHIC software. The lower number of windows using the SNP-based approach than in the physical size approach (11 compared to 25 per recombination rate region, respectively) is expected to increase baseline accuracies for the 1-in-1 accuracy method, as these accuracies constitute proportions of the most extreme values observed.

We used a total of 16 population genetic summary statistics grouped into five families based on the type of genomic information they use. The five families are 1) metrics based on the site frequency spectrum (SFS): nucleotide diversity π (Tajima 1983), the difference in nucleotide diversity between the two post-divergence populations $\Delta\pi$, $\hat{\theta}_W$ (Watterson 1975), and *Tajima's D* (Tajima 1989); 2) metrics based on genotype distance: the variance, skewness, and kurtosis of the genotype distance distribution g_{kl} (Kern and

Schrider 2018); 3) metrics based on the distribution of multilocus genotype frequencies: $nDiplos$ (the number of distinct genotype vectors), J_1 , J_{12} , and J_2/J_1 (or G_1 , G_{12} , and G_2/G_1 , the multilocus genotype equivalents of the haplotype-based metrics H_1 , H_{12} , H_2/H_1 ; Messer and Petrov 2013; Garud et al. 2015; Harris et al. 2018); 4) metrics based on linkage disequilibrium (LD): $diploZ_{NS}$ and $diplo\omega$ (the genotypic LD equivalents of Z_{NS} (Kelly 1997) and ω (Kim and Nielsen 2004)) as well as $H - scan$ (Schlamp et al. 2016); and 5) measures of population differentiation: F_{ST} (Wright 1949) and the genetic beta diversity measure $\bar{B}^{GD}$ (Reeve et al. 2016; White 2019; described below).

Metrics based on the SFS (family 1) make use of the distribution of allele frequencies in the genome. Nucleotide diversity π is defined as the number of nucleotide differences per site between two randomly chosen sequences from a population (Tajima 1983). The Watterson estimator $\hat{\theta}_W$ measures the population mutation rate, which is the product of the effective population size and the mutation rate (Watterson 1975). *Tajima's D* is defined as the average number of pairwise differences divided by the number of segregating sites (Tajima 1989). A negative *Tajima's D* indicates an excess of rare alleles, which can be caused by a recent selective sweep. In metrics based on genotype distance (family 2), the genotype distance distribution g_{kl} between individuals k and l is defined as $g_{kl} = \sum_{i=0}^m 1_{x_{ki} \neq x_{li}}$, which is the sum of unequal genotypes in a genotype vector x_k (Kern and Schrider 2018), where the multilocus genotype $x_k = \{x_{k0}, x_{k1}, \dots, x_{km}\}$ denotes the set of genotypes in individual k . In metrics based on the distribution of multilocus genotype frequencies (family 3), $nDiplos$ is simply the number of distinct multilocus genotypes in a population. J_1 is the multilocus genotype homozygosity calculated using all multilocus genotypes. J_{12} is the multilocus genotype homozygosity calculated using the two most frequent multilocus genotypes. J_2/J_1 is the ratio of multilocus genotype homozygosity calculated using all but the most frequent multilocus genotype to J_1 (Messer and Petrov 2013; Garud et al. 2015; Harris et al. 2018). Metrics based on LD (family 4) depend on the association of alleles at different loci in a population. $diploZ_{NS}$ uses multilocus genotypes to estimate the squared correlation of alleles at pairs of polymorphic sites (Kelly 1997; Kern and Schrider 2018). $diplo\omega$ is a related metric that is based on the sum of correlation coefficients between sets of polymorphic sites (Kim and Nielsen 2004), which can also be calculated from multilocus genotypes (Kern and Schrider 2018). $H - scan$ is a method that aims to detect selective sweep signatures by measuring the average length of pairwise homozygosity tracts along the genome (Schlamp et al. 2016). The differentiation-based metric F_{ST} (family 5) measures the between-population variance in allele frequencies (Wright 1949).

All metrics except $H - scan$ and population differentiation (F_{ST} and $\bar{B}^{GD}$) were estimated using diploS/HIC on ms files for populations 2 and 3 separately and constitute standardised values (as described in Kern and Schrider 2018). For diploS/HIC metrics and $H - scan$, only the results for the population experiencing positive selection (population 2) are shown. $H - scan$ was calculated using the software provided with the original publication (Schlamp et al. 2016). F_{ST} and $\bar{B}^{GD}$ were calculated between populations 2 and 3 based on vcf files. F_{ST} was estimated using vcftools v0.1.17 (Danecek et al. 2011). $\bar{B}^{GD}$ was estimated using the R package rdiversity v2.0 (Reeve et al. 2016; Mitchell, Reeve, and White 2020).

3.3.3 $\bar{B}^{GD}$, a beta diversity measure using genetic distances derived from information theory

$\bar{B}^{GD}$ is a measure of between-population diversity (i.e. beta diversity) based on genetic or genomic data (White 2019). $\bar{B}^{GD}$ is calculated from the similarity between individuals (see Reeve et al. 2016, Tables 1 and 2 for further details) and is mathematically independent from within-population diversity (Jost 2010). The similarity between individuals is defined as $z_{ij} = 1 - k \times n_{ij}$, where n_{ij} is the number of differences in sequence between individuals i and j in the genomic window being considered (with homozygotes for the opposite allele having 2 changes and heterozygotes compared to homozygotes having 1 change), and k is a scaling constant less than the inverse of the maximum number of changes observed (so $z_{ij} \geq 0$). In trials, the scaling factor $1/k$ has been set to 100, 1000, or was automatically determined by the maximum distance values. We found a value of $k = 1/1000$ to lead to the highest accuracies in sweep prediction for our data set. The resultant measure, $\bar{B}^{GD}$, derived from $\bar{B}$, the effective number of distinct populations (or communities) (Reeve et al 2016), measures the number of genetically distinct genomic windows (GD = Genetic Distance), which takes a value of 1 when the windows have exactly the same genotypes in both populations (i.e. no differentiation), and a maximum value of 2 when the windows are completely genetically distinct. A parameter q , $0 \leq q \leq \infty$, determines the relative importance of rare and common sequences in the analysis (see Hill (1973) and Reeve et al. (2016) for further details on q). Three values of q were tested (0, 1, and 2). Since the choice of q had a very minor effect on the results, we only report $\bar{B}^{GD}$ for $q = 2$ in the main text.

3.3.4 Accuracies of sweep prediction

To determine how well each metric predicted the presence and location of mutations under selective sweeps, accuracies were calculated on the outcome across the 100 replicates in the following ways: 1) '1-in-1 window accuracy': the proportion of replicates in which the window containing the selected mutation (the central window) had the most extreme value (peak or valley/highest or lowest, depending on the metric) in each recombination region; 2) '1-in-3 windows accuracy': the proportion of replicates in which the window containing the selected mutation or its neighbouring one up- and one down-stream windows had the most extreme value, which is designed to also capture less distinct signals of selection. These accuracies therefore represent the proportion of simulation replicates with a summary statistic peak (or valley) in the central window of a recombination region, which was found to be a useful proxy for the observation of summary statistic peaks in the present study. We only report results based on the 1-in-1 window accuracy in the main manuscript, since we found that the relative performance of metrics was essentially the same using the 1-in-3 windows accuracy. For the analyses based on windows holding 21 SNPs, we only calculated 1-in-1 window accuracies (Figures S4–S7). (In the online supplementary material of the original publication we also describe and report results from an accuracy metric based on comparisons of simulations with selection to neutral simulations, which are comparable to the proportion of simulation replicates with P values more extreme than in simulations of neutral scenarios.) All analyses of accuracy were performed using R v3.5.1 (R Core Team 2018). R scripts, simulated genetic data, and SLiM scripts are deposited at <http://dx.doi.org/10.5525/gla.researchdata.981>.

3.4 Results

3.4.1 Patterns of selective sweeps across the genome

When averaging the summary statistics across replicates of each scenario, most metrics showed relatively clear peaks at or around the selected mutations (see Figure 3.2 for an example; see online supplementary material of original publication for plots of all scenarios). These peaks were more distinct in the differentiation measures ($\bar{B}^{GD}$ and F_{ST}) compared to the other metrics used. Measures based on genotype distance (variance, skewness, and kurtosis of g_{kl}), the multilocus genotype frequency metrics J_1 and J_{12} as well as $H - scan$ also frequently had distinct peaks, while other measures based on multilocus genotype frequencies (J_2/J_1 and $nDiplos$) and LD (in particular $diploZ_{NS}$ and $diplo\omega$) were more prone to noise, with less distinct peaks around the selected mutations

(Figure 3.2). The distinctness of peaks around the genomic region under selection is directly reflected in the sweep prediction accuracies reported in more detail below.

3.4.2 Accuracy in identifying sweep location across scenarios

The ability of the 16 different summary statistics to accurately infer selective sweeps varied dramatically across scenarios. Generally, the more distinct the peaks observed in the graphs were (Figure 3.2), the higher the accuracy in predicting the true location of the selective sweep. Figure 3.3 shows results obtained for all four types of scenarios (sympatric speciation and three secondary contact scenarios) and differing population growth and migration parameters, under hard and soft sweep simulations, in the region with standard recombination rate in simulations with equal population sizes. $\bar{B}^{GD}$ outperformed all other metrics in finding the correct location of the mutation under selection. The mean accuracy of F_{ST} was, on average, slightly lower (Figure 3.3, Figures S3.1–S3.3). Of the diploS/HIC selection metrics, genotype distance-based metrics performed best, especially the skewness and kurtosis of g_{kl} . Of the metrics based on the SFS, only $\Delta\pi$ had comparable accuracies to g_{kl} metrics. $H - scan$ had somewhat higher accuracies compared to the SFS-based $\hat{\theta}_W$ and other metrics based on LD, which had the lowest accuracies. Metrics based on multilocus genotype frequencies performed better than most LD metrics but not as well as the g_{kl} metrics. Additional results are given in appended figures for the low recombination rate region in simulations with equal population sizes and unequal population sizes (Figures S3.1 and S3.2, respectively) and standard recombination rate region simulations with unequal population sizes (Figure S3.3). Results for the windows based on holding 21 SNPs mostly agreed with those based on physical size of windows and are given in Figure S3.4 for the standard recombination rate region with equal population sizes, Figure S3.5 for the standard recombination rate region with unequal population sizes, Figure S3.6 for the low recombination rate region with equal population sizes, and Figure S3.7 for the low recombination rate region with unequal population sizes. The SNP-based window approach led to overall higher accuracies compared to the physical window approach due to the lower number of windows used (11 compared to 25 per recombination rate region, respectively). The accuracies increased more in some metrics than in others, particularly in the case of π and the metrics based on multilocus genotype frequencies. However the overall pattern of performance was consistent to the window approach.

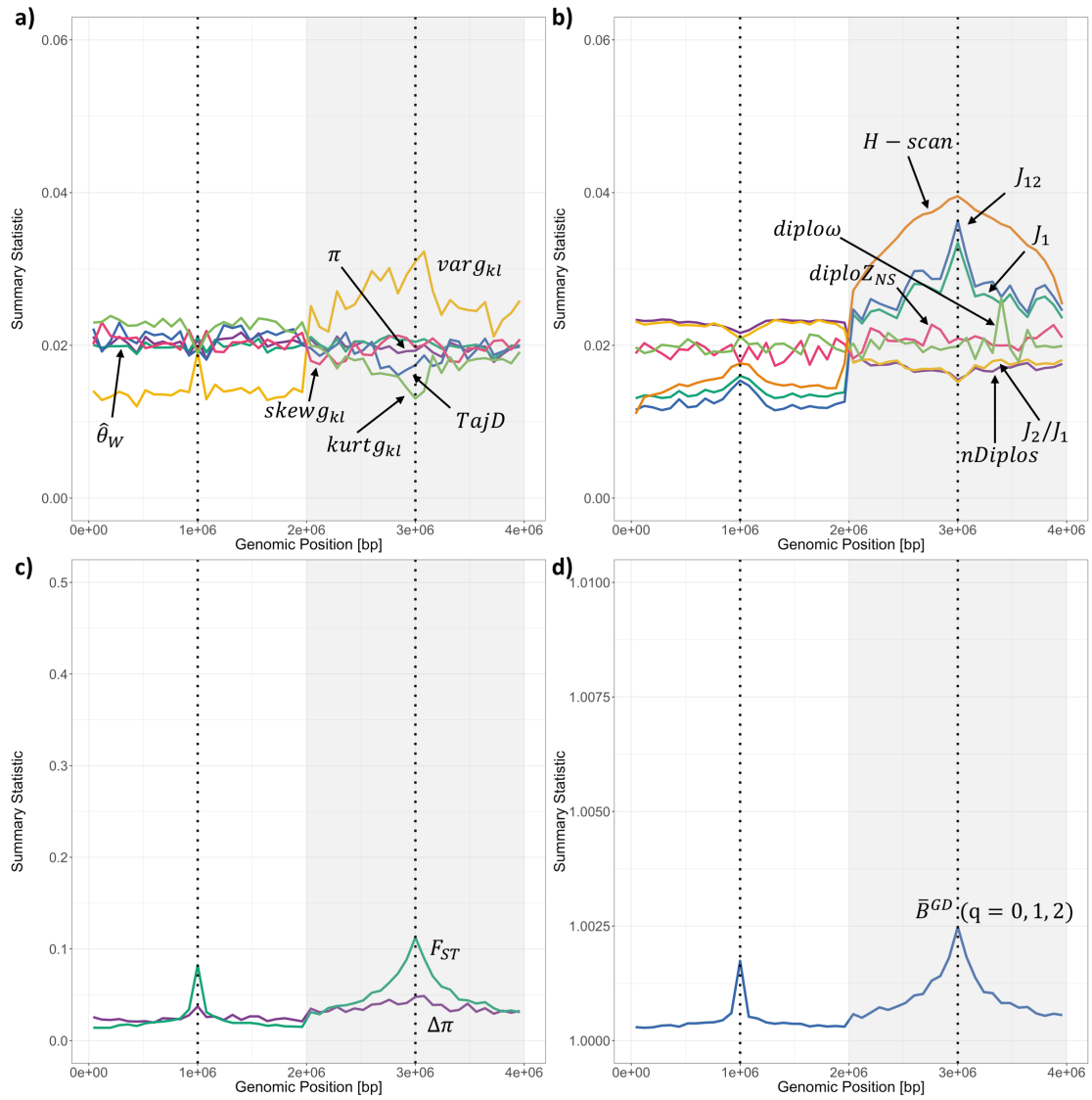


Figure 3.2 – Traces of summary statistics for an example simulation scenario (average across 100 replicates): Scenario 2 (secondary contact), with hard sweeps and constant, unequal population sizes with a migration rate of 0.01. Summary statistics a) $var g_{kl}$ and $kurt g_{kl}$ show clear peaks centred around the locations of the selective sweeps (dotted lines). b) J_1 , J_{12} , and $H - scan$ show relatively clear peaks around the sites of the selected mutations. c) $\Delta \pi$ varies across the genome and shows only subtle peaks, while the signals of F_{ST} around the mutation under selection are distinct. d) $\bar{B}^{GD}$ shows distinct peaks around the sweep location. The left part of each panel is the standard recombination region (white background) while the right part is the low recombination region (grey background). The graphs shown are derived by averaging across 100 simulation replicates. For plotting, the absolute values of $\Delta \pi$ multiplied by 10 and the values of $H - scan$ divided by 1,000,000 were used. $\bar{B}^{GD}$ is shown for $q = 0$, $q = 1$, and $q = 2$, but the graphs are, however, indistinguishable due to almost identical $\bar{B}^{GD}$ values.

3.4.3 Prediction of sweep location - scenario-dependent performance

The prediction accuracies of demographic scenarios depended substantially on migration rate and the metric. For low migration rates ($m = 0.001$), all four demographic scenarios had similar prediction accuracies (Figure 3.3; Figures S3.1–S3.7). For intermediate migration rates ($m = 0.01$), secondary contact scenarios with intermediate to strong negative selection in population 3 (scenarios 3 and 4) had the highest prediction accuracies. This was most pronounced for multilocus genotype frequency-based metrics, while differentiation-based metrics and $\Delta\pi$ had more similar prediction accuracies across scenarios in most cases. For high migration rates ($m = 0.1$), the secondary contact scenario with the strongest negative selection in population 3 (scenario 4) had the lowest prediction accuracies across metrics (Figure 3.3; Figures S3.1–S3.7).

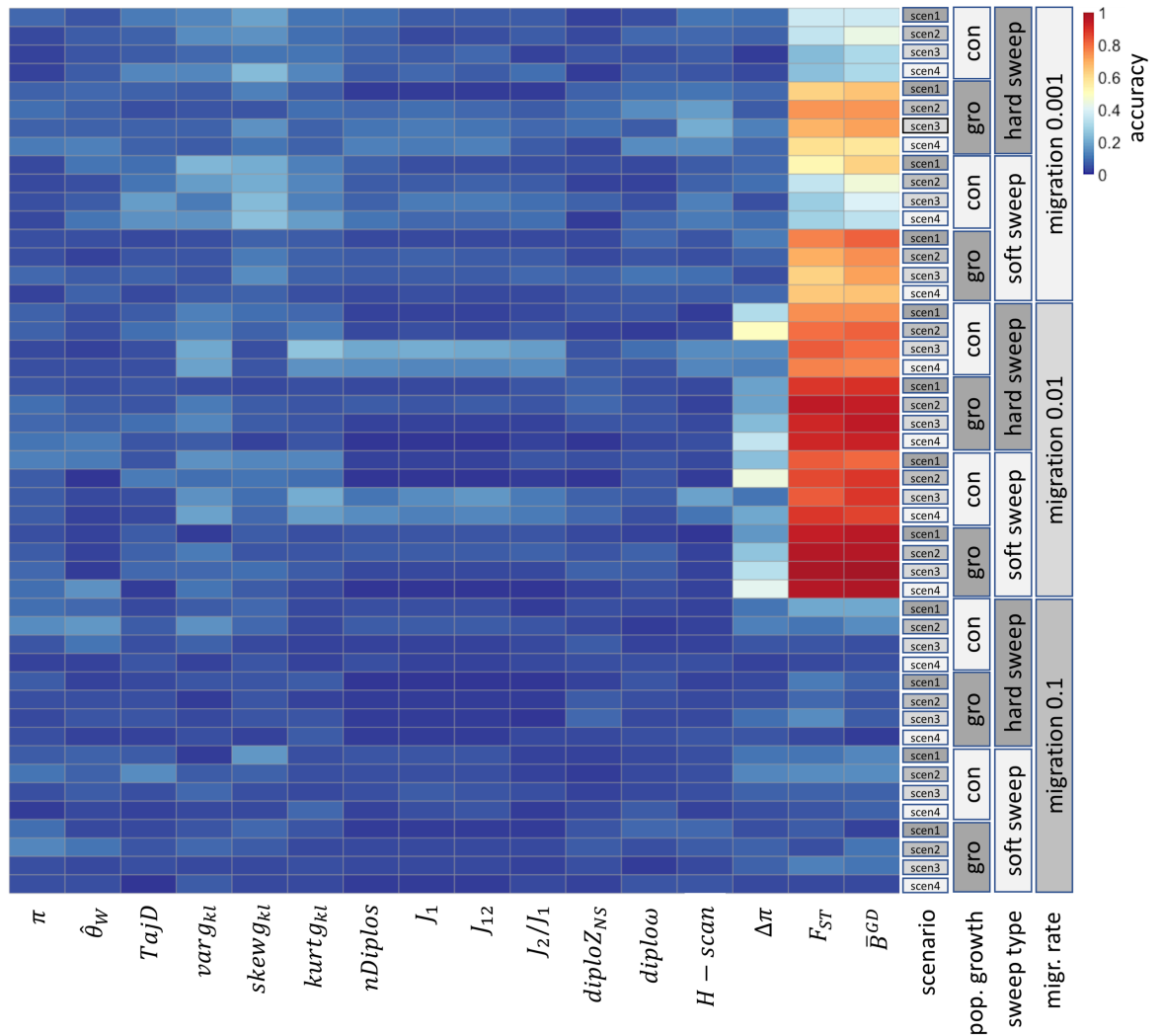


Figure 3.3 – A heat map of prediction accuracies of the location of selective sweeps for the standard recombination rate region and equal population sizes. The colours in the heat map are based on accuracy values (see scale at the top right of the plot, with the highest accuracies, 1, in dark red and the lowest accuracies, 0, in dark blue), which are the proportions of 100 simulation replicates in which the correct window was identified as

containing the selected mutation (1-in-1 window approach). For $\bar{B}^{GD}$, the chosen q value was 2. scen1 = scenario 1 (sympatric divergence); scen2 = scenario 2 (secondary contact scenario with low negative selection in population 3); scen3 = scenario 3 (secondary contact scenario with intermediate negative selection in population 3); scen4 = scenario 4 (secondary contact scenario with strong negative selection in population 3); gro = logistically increasing population size; con = constant population size; soft sweep = soft selective sweep scenario; hard sweep = hard selective sweep scenario; migration = migration rate between populations 2 and 3; pop. growth = population growth (gro or con); migr. rate = migration rate.

Suggested steps for locating selective sweeps

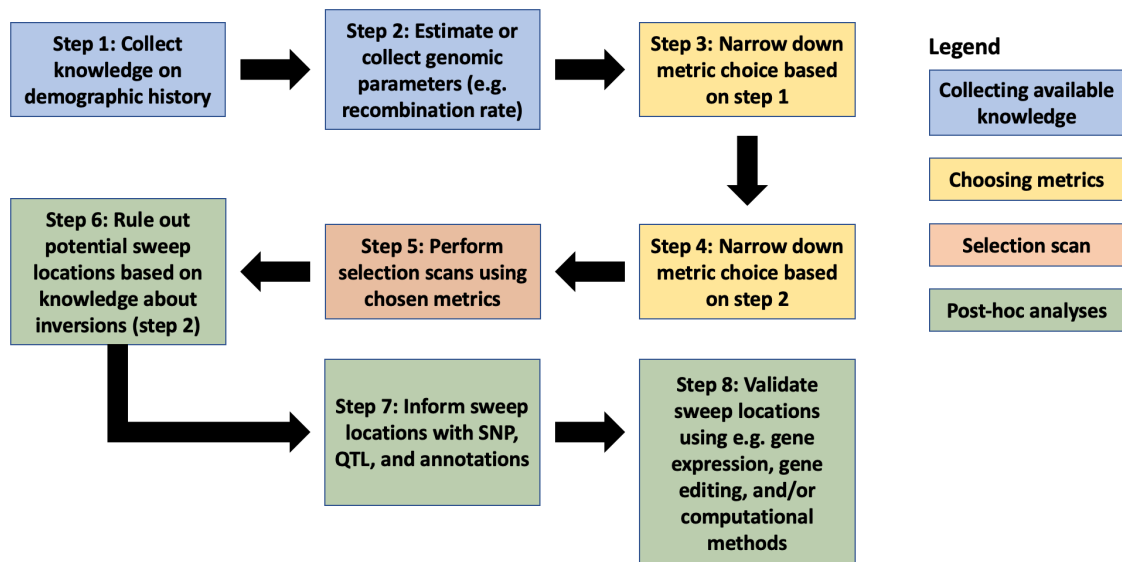


Figure 3.4 – Suggested steps for detecting the genomic position of selective sweeps.

Drawing from the simulation results, empirical studies as well as theory, we suggest an eight-step process that can help to locate and make sense of selective sweeps in whole-genome data.

Of all parameters, migration rate had the strongest influence on accuracies of sweep location prediction. Intermediate migration rates ($m = 0.01$) between the two diverging populations led to higher accuracies of sweep prediction for differentiation-based metrics and $\Delta\pi$, with low migration rates ($m = 0.001$) having intermediate accuracies, and high migration rates ($m = 0.1$) having the lowest accuracies (Figure 3.3; Figures S3.1–S3.7). However, the skewness of g_{kl} , $Tajima's D$, and $H - scan$ performed best at low migration rates, while the variance and kurtosis of g_{kl} as well as multilocus genotype frequency-based metrics performed equally well at low and intermediate migration rates.

For almost all metrics and scenarios, highest migration rates led to the lowest accuracies, especially in the case of secondary contact scenarios with intermediate to strong negative selection in population 3 (scenarios 3 and 4). Differences in migration rate had the strongest effect on differentiation-based measures (F_{ST} and $\bar{B}^{GD}$). At high migration rates, F_{ST} and $\bar{B}^{GD}$ had similarly low or only slightly higher accuracies compared to the other metrics. Accuracies were also similar in genotype distance-based and differentiation-based metrics at low migration rates and constant population sizes.

The influence of recombination rate on sweep prediction accuracies was comparably high to that of migration rate. Across all metrics, the accuracies were substantially higher in the standard recombination rate region compared to the low recombination rate region (Figure 3.3; Figures S3.1–S3.7). Moreover, for low recombination rates, secondary contact scenarios had lower prediction accuracies than sympatric divergence scenarios, while this was not true for standard recombination rates. For measures based on differentiation and those based on LD, there was barely any difference between the accuracies of sweep prediction in hard sweep scenarios compared to soft sweep scenarios. However, the influence of sweep type was greater in most other metrics, with higher prediction accuracies in the case of hard sweeps especially in multilocus genotype frequency-based metrics.

Scenarios that included demographic growth led to higher accuracies in sweep prediction across metrics (Figure 3.3; Figures S3.1–S3.7). Measures based on multilocus genotype frequencies and $\Delta\pi$ had higher prediction accuracies in scenarios with a larger population 3 (the population with negative selection), while for other metrics the prediction accuracies were mostly unchanged (Figure 3.3; Figures S3.1–S3.7).

3.5 Discussion

3.5.1 The pros and cons of different metrics

Our results demonstrated that different summary statistics varied dramatically in their ability to infer simulated selective sweeps under the demographic conditions we evaluated. Differentiation-based measures F_{ST} and particularly the beta diversity measure $\bar{B}^{GD}$, which was evaluated robustly for the first time here, outperformed metrics based on the SFS (π , $\Delta\pi$, $\hat{\theta}_W$, and *Tajima's D*), genotype distance (the variance, skewness, and kurtosis of the genotype distance distribution g_{kl}), multilocus genotype frequencies (*nDiplos*, J_1 , J_{12} , and

J_2/J_1), and LD ($diploZ_{NS}$, $diplo\omega$, and $H - scan$) across the majority of evolutionary and demographic scenarios.

Accuracies based on identifying the correct genomic location where the selective sweep occurred were, on average across scenarios, highest in $\bar{B}^{GD}$ and second highest in F_{ST} , with the best performance at intermediate and the lowest at high migration rates. This was true across sweep types, population sizes, and sympatric and all secondary contact scenarios (i.e. scenarios 1 to 4). This is also evident in the observed distinctness of the peaks in these metrics centred around the selected mutations in a large number of scenarios. Of the metrics that are not based on regions of genomic differentiation, relatively high accuracy in predicting selective sweeps was also found in $\Delta\pi$ and those derived from the distribution of genotype distance. Multilocus genotype frequency-based metrics and $\Delta\pi$ performed best at intermediate migration rates but poorly at other migration rates. In some scenarios, especially at low migration rates, genotype distance-based metrics such as the skewness of g_{kl} performed similarly well as F_{ST} and $\bar{B}^{GD}$. High performance of genotype distance-based metrics (for both hard and soft sweeps) and multilocus genotype frequency metrics (for hard sweeps) has also been observed in other simulation studies and using empirical data (Kern and Schrider 2018). Using SNP-based windows, we observed a slight relative increase in performance for the metrics π and those based on multilocus genotype frequencies, while other metrics were less affected by how windows had been defined. The lowest accuracies were observed using LD-based metrics, especially $diploZ_{NS}$ and $diplo\omega$, which is in concordance with the observation of rather poor performance compared to methods based on allele frequency changes (i.e. SFS and F_{ST}) (Kim and Nielsen 2004). However, other methods based on LD, most of which require phased data, could have higher power (Vatsiou et al. 2016).

3.5.2 The influence of demography, recombination, and sweep type

Our simulations show that gene flow influences how accurately the genomic location of selective sweeps can be predicted in a non-linear way, as both low and high migration rates (low or high m) led to lower prediction accuracies for many metrics. In the case of low migration rates, this is likely a result of increased noise due to increased overall genome-wide differentiation between populations (Ravinet et al. 2017). On the other hand, high migration rates can lead to the breakdown or slowdown of selective sweeps, even more so in scenarios with strong negative selection in the population without selective sweep (scenario 4). Therefore, in study systems with high ongoing gene flow, as has been observed in recent speciation events in the apple maggot *Rhagoletis pomonella* and in

Midas cichlids (*Amphilophus* sp.) (Elmer et al. 2014; Egan et al. 2015; Kautt et al. 2016), the potential for inferring the presence and location of selective sweeps might be reduced and depends on the strength of selection and the presence of negative selection. Secondary contact scenarios, which all included an allopatric phase with zero migration before resuming gene flow after secondary contact, decreased the accuracies of predicting the sweep location when the recombination rate was low. These observations imply that in study systems that experienced secondary contacts, such as has been reported for *Cottus* sp. and *Mimulus* sp. (Nolte et al. 2006; Kenney and Sweigart 2016; Lucek et al. 2018), screens for selection will have reduced power in regions of the genome with reduced recombination. On the other hand, in systems with good evidence for sympatric speciation, as for instance in Midas cichlids, East African crater lake cichlids, and *Howea* palms (Savolainen et al. 2006; Malinsky et al. 2015; Kautt et al. 2016; Torres-Dowdall et al. 2017; Osborne et al. 2019; Papadopulos et al. 2019), screens for selective sweeps in genomic data will have increased power to detect signals of selection in low-recombination regions of the genome. The detection of selective sweeps in genomic regions with standard recombination rates, however, depends more on the particular combination of demographic history, migration rate, and the chosen metric. This shows that, ideally, determining the demographic history of a population, the recombination landscape, and the amount of gene flow should precede selection scans, which is particularly important in systems where the demographic history varies strongly from population to population (e.g. Jacobs et al. 2020).

Accuracies were lower in the low recombination rate region compared to the standard recombination rate region. This is consistent with the observation of false positive signals of selection caused by the decreased local recombination rate due to genomic rearrangements (e.g. inversions) or other influences on genome organisation and structure (Thornton 2005; Lotterhos 2019). A higher number of false positives should therefore also be expected in taxa with overall reduced recombination rates, as for example in opossum *Monodelphis domestica* and in *Oenothera* sp. (Mikkelsen et al. 2007; Rauwolf et al. 2011). The negative effect of lower recombination rate was similarly strong for all metrics, which is to be expected given that all of these metrics are more or less based on a reduction in local diversity around the mutation under selection or the association of alleles (Pritchard et al. 2010; Messer and Petrov 2013).

As noted by Kern and Schrider (2018), metrics based on multilocus genotype frequencies are expected to have lower power to detect soft selective sweeps compared to hard selective sweeps, which is also what we observed here. This effect was somewhat smaller or was not observed at all for other metrics. Differentiation-based metrics in

particular, showed relatively similar prediction accuracies for hard and soft sweeps. This is in concordance with various empirical and theoretical studies based on genome-scale data (Chen et al. 2010; Hohenlohe et al. 2010; Vitti et al. 2013; Pavlidis and Alachiotis 2017). The fact that a substantial number – and in many cases the majority – of sweeps constitute soft sweeps underlines the importance of using metrics that are robust to identifying regions under selection in both hard and soft sweeps (Pritchard et al. 2010; Messer and Petrov 2013; Schrider and Kern 2017). That some metrics are unaffected by soft selective sweeps while other metrics are influenced by both hard and soft sweeps could itself potentially be used to distinguish sweep type. Garud et al. (2015) and Harris et al. (2018) showed that the joint use of the metrics J_{12} and J_2/J_1 (called G_{12} and G_2/G_1 in Harris et al. 2018) enables the detection of and discrimination between hard and soft selective sweeps, which is reflected in our results to some extent. Machine learning approaches such as diploS/HIC (Kern and Schrider 2018), which makes use of the J_{12} and J_2/J_1 metrics as well, are also promising tools for the discrimination between signatures of hard and soft sweeps.

3.5.3 The utility of differentiation-based metrics

Differentiation-based metrics have found wide application in screens for positive selection (Chen et al. 2010; Narum and Hess 2011; De Villemereuil and Gaggiotti 2015). Despite its popularity and good performance in a range of situations, the use of F_{ST} has received criticism as well (Bierne et al. 2013; Cruickshank and Hahn 2014; Lotterhos and Whitlock 2014). The drawbacks of F_{ST} have been one of the motivations for the application of $\bar{B}$ to genetic data, as $\bar{B}^{GD}$ (White 2019). The better performance of $\bar{B}^{GD}$ compared to F_{ST} likely derives from two desirable properties that $\bar{B}^{GD}$ has but F_{ST} does not have. Firstly, $\bar{B}^{GD}$ was observed to be monotonically increasing with increasing time or strength of divergence (White 2019). F_{ST} , on the other hand, is not monotonic with increasing divergence time and strength (Wang 2012; White 2019). Secondly, $\bar{B}$ has the property that low within-population diversity does not have an effect on the magnitude of $\bar{B}$ (Reeve et al. 2016), or that alpha diversity is mathematically independent from beta diversity (Jost 2010). Rather, only between-population diversity or differentiation affects the magnitude of $\bar{B}$. F_{ST} , however, is substantially affected by low within-population diversity between the two populations, potentially leading to false positive differentiation outliers, as has been shown for a number of studies (Cruickshank and Hahn 2014). These two properties should make $\bar{B}^{GD}$ the better choice if one wants to truly measure differentiation between populations, although further model-based and empirical assessments will be needed.

However, despite the good performance of differentiation-based measures, which could be expected given they *de facto* constitute comparisons to a reference, which is not true for the other (single-population) metrics, differentiation-based metrics are not always appropriate. In particular, $\bar{B}^{GD}$ and F_{ST} as well as $\Delta\pi$ require at least two (related) populations, one of which should be a reference population that does not experience selection at the same genomic location. This means that the signatures of selection will be lost if the selected mutation is also selected for in the reference population (i.e. in the case of global selection). In fact, such cases can even lead to reduced differentiation in the selected regions compared to neutral regions, which would be even more pronounced in low-migration scenarios (Zheng and Wiehe 2019). It is also possible for beneficial mutations in one population to be neutral in the other population, even though such cases might be less common than previously thought (Kern and Hahn 2018). In many study systems such a reference population, for example an ancestral population in a population pair from which a derived population diverges, is not available. An ideal scenario would be a population pair under divergence with distinct ecological preferences that lead to adaptation driven by differential selection. Examples for such a scenario are the divergences of benthic and pelagic ecomorphs in three-spined stickleback (*Gasterosteus aculeatus*) (Schluter and McPhail 1992; Bolnick and Lau 2008) or Arctic charr (*Salvelinus alpinus*) (Elmer 2016; Jacobs et al. 2020). Another potential drawback of differentiation-based metrics is the fact that it has been shown that hidden or hierarchical population structure can negatively affect the detection of natural selection based on differentiation measures (Excoffier et al. 2009). This implies that clear knowledge about the population structure at each level is required for sound application of differentiation-based metrics. It has also been pointed out that differences in sample size among populations, the dependency on minor allele frequencies, and inconsistencies among estimators can decrease the reliability of F_{ST} (Berner 2019), which is why allele frequency difference (AFD) has been proposed as an alternative. We note that a detailed assessment of the effects of background selection, which has been shown to affect estimates of F_{ST} (Matthey-Doret and Whitlock 2019), and other confounding factors on the detection of selective sweeps, especially using $\bar{B}^{GD}$, should be performed in future simulation studies.

3.5.4 Potential caveats

In the current study, we did not account for some factors of empirical data that could potentially affect the performance of summary statistics to detect selective sweeps. This includes poor data quality, insufficient coverage, and inaccuracies (e.g. erroneously called

SNPs), which have been shown to negatively impact population genetic estimates (Crawford and Lazzaro 2012).

A number of metrics were not included in this simulation study. Those excluded mostly were ones that belong to the same families of metrics that we included here, i.e. are based on or depend on the SFS, genotype distance, multilocus genotype frequencies, LD, or population differentiation. The decision only to include metrics that do not require phased data or a polarised SFS stems from the fact that phased data and ancestral sequences are not readily available and/or are difficult to obtain for most empirical studies, especially in non-model ecological and evolutionary study systems. The results presented in this paper highlight the importance of knowing the properties of different metrics under different demographic scenarios.

3.5.5 Suggested steps for locating selective sweeps

Drawing from our simulation results, empirical studies, and theory, we suggest an eight-step process that can help to locate and make sense of selective sweeps in whole-genome data (Figure 3.4):

- 1) Collect knowledge on the underlying demographic history of the studied populations to create a basis for subsequent analyses.
- 2) Genomic properties and parameters (i.e. recombination rate or the location of inversions) of a study system can often be estimated using available data (and in some cases even the same kind of data) to further inform downstream analyses.
- 3) Based on the knowledge gained in 1), choose a handful of metrics that proved suitable for a given demographic scenario. One of the most important observations from our exploration is that many metrics such as multilocus genotype frequency-based metrics and $\Delta\pi$ are most powerful at intermediate migration rates while genotype distance-based and differentiation-based metrics show a more consistent performance across migration rates. A combination of genotype distance-, multilocus genotype frequency-, differentiation-based metrics, and $\Delta\pi$ could be a powerful way of screening for selection in unphased genomic data. It should also be noted that, given the low accuracies of identifying the correct sweep location of many metrics under various demographic scenarios, complete overlap of candidate genomic regions under selection inferred using different methods can be considered unlikely. However, even partial overlap of candidate regions across metrics is stronger evidence for a selective sweep than that based on only one metric. Since metrics of the same family (e.g. metrics based on multilocus genotype frequencies) showed a correlated or collinear response to selection under a given demographic scenario, metrics

in empirical studies should rather be chosen from multiple metric families that have proven to be powerful for a given scenario. In this context, the development of composite measures of selection that capture the combined signatures of different metrics (Lotterhos et al. 2017; Verity et al. 2017) will be an important future avenue for detecting selective sweeps.

4) If available, recombination rate estimates across the genome from 2) can further narrow down the choice of metrics for screening for selective sweeps. As this is not feasible in many cases, taking into account global recombination rate estimates of the same or related species can be useful when choosing suitable metrics. An important result is that, at low recombination rates, *H* – *scan* and multilocus genotype frequency-based metrics were more powerful than most other single-population metrics and could complement differentiation-based and genotype distance-based metrics.

5) Perform selection scans using the chosen population genetic metrics.

6) When potential locations of selective sweeps are identified, knowledge from 2) can be used to rule out factors that could lead to false positive signals of selection (e.g. inversions or other genomic rearrangements). However, the presence of inversions, for example, cannot rule out selective sweeps or other forms of selection. In fact, it has been shown in several study systems, from Atlantic cod (*Gadus morhua*) and rough periwinkles (*Littorina saxatilis*) to seep monkeyflowers (*Mimulus guttatus*), that inversions can harbour genes with high importance for adaptation (Kirubakaran et al. 2016; Puzey et al. 2017; Morales et al. 2019; Refoyo-Martínez et al. 2019). In many cases, however, the location of inversions cannot be determined due to insufficient and/or unsuitable data.

7) The candidate genomic regions can then be analysed and explored, for example with regard to the functional genes or regulatory regions they contain. Sets of outlier genes in candidate regions can be analysed for gene ontology term or similar functional enrichment (Gaudet and Dessimoz 2017). Using available genome information, SNPs in these regions can be annotated, for example with software such as SnpEff (Cingolani et al. 2012). It can also be helpful to analyse allelic shifts of these SNPs in more detail (potentially also from other populations or related study systems). If available, candidate sweep regions could be analysed for overlap with quantitative trait loci in the same or related taxa (Stinchcombe and Hoekstra 2008; Jacobs et al. 2017).

8) Ultimately, while not currently feasible for most ecological and evolutionary non-model organisms, the last step could involve validating candidate sweep regions, genes, and alleles, for example using analyses of gene expression (Rahi et al. 2017; McGirr and Martin 2018), gene editing and knockouts (Fernandez i Marti et al. 2018; Chao et al.

2019), or computational methods for predicting functional changes such as modified enzyme function or protein structure (Tang and Thomas 2016; Lee et al. 2017).

Our results demonstrate that, if the conditions are met, differentiation-based approaches are a powerful way to detect selective sweeps in whole-genome data. In particular the measure $\bar{B}^{GD}$ proposed here, a beta diversity measure derived from information theory principles, proved to outcompete the other metrics used, making it a promising method for screening for selective sweeps and differentiation. Lastly, other recently described metrics (Kern and Schrider 2018), mostly those based on genotype distance, were shown here to be the most powerful non-differentiation metrics for screening for selective sweeps in unphased whole-genome data.

Chapter 4: Shared genomic targets of selection in repeated sympatric divergences of Arctic charr

4.1 Abstract

The salmonid fish the Arctic charr (*Salvelinus alpinus*) is a highly variable species, exhibiting repeated cases of parallel sympatric divergence into strongly phenotypically and genetically differentiated ecomorphs in numerous northern freshwater lakes. Despite being a model for ecological speciation, the genetic components under shared selection in these parallel ecomorph divergences have not been obtained. To identify the functional genomic targets of selection and the extent to which they are shared across repeated independent divergences, we sequenced the whole genomes of six populations of Arctic charr including five benthic-pelagic pairs. We recovered a robust set of shared genes under selection in all replicates; these have functional roles in other organisms and lie in genomic regions of known ecologically relevant QTLs for salmonids. The shared genes under selection did not contain any substitutions in the coding regions but only in the up- and downstream regions of genes. Melanosome-related processes and gene regulation were shared targets of selection across populations. A large proportion of variation was population-specific, reflecting that the ecomorphological parallelism occurs despite differing genetic histories and patterns within lakes. The combination of population genetic metrics and overlap with QTLs we used, together with the sharing of outliers across populations, are strong indicators of a common role of the obtained outlier genes in ecological adaptation of Arctic charr.

4.2 Introduction

One of the most pressing questions in the study of evolution is what role contingency (as opposed to deterministic forces such as adaptation by natural selection) plays in the evolutionary process (Stern and Orgogozo 2009; De Visser and Krug 2014; Losos 2017; Langerhans 2018; Nosil et al. 2018). In nature, predictability of the process of adaptation can be studied by looking at the repeatability, convergence, or parallelism of local adaptation, especially in the context of replicated ecological divergence (Elmer and Meyer 2011; Conte et al. 2012; Jones et al. 2012; Kowalko et al. 2013; Mahler et al. 2013; Elmer et al. 2014; Oke et al. 2017; Stuart et al. 2017; Bolnick et al. 2018; Nosil et al. 2018; Miller et al. 2019; Jacobs et al. 2020). Arctic charr (*Salvelinus alpinus*) are characterised by such repeated cases of divergence into strongly phenotypically differentiated ecomorphs that

most often differentiate along the benthic-pelagic axis in freshwater lakes across their Holarctic distribution. Moreover, Arctic charr exhibit exceptionally high potential for diverging and diversifying rapidly. For this reason, charr has become a model system for the study of ecological adaptation (Klemetsen 2013; Elmer 2016; Jacobs et al. 2020).

This high potential for ecological divergence in Arctic charr has long been attributed to as-yet undetermined genetic differences and a contribution of phenotypic plasticity (Skúlason et al. 1996; Adams et al. 2004; Alexander and Adams 2004). Research based on populations across Eurasia suggested that at least some of the repeated phenotypic divergence is attributable to genetic influences on gene expression (parallel *cis*-eQTL; Jacobs et al. 2020), with gene expression showing significant parallelism and only very few loci (from reduced representation sequencing) exhibiting parallel differentiation across populations (Jacobs et al. 2020). Nonetheless, in this previous study, hundreds of loci were significantly associated with ecomorph divergence and there was a strong correlation between the extent of genetic and phenotypic differentiation in sympatry (Jacobs et al. 2020). Combined with the evidence of parallelism in gene regulatory networks and *cis*-eQTL, this is a strong indication that specific between-population genetic changes can be, but do not have to be, parallel in order for gene expression and phenotypic changes to be parallel. High levels of phenotypic parallelism but low levels of genome-wide parallelism have also been shown in other systems, such as three-spined stickleback (*Gasterosteus aculeatus*) (Fang et al. 2020; Kirch et al. 2020), where despite high levels of parallel divergence in phenotype, genetic parallelism is restricted to differentiation of only 0.1% of the genome and a small set of shared ancestral genetic variation (Jones et al. 2012; Roesti et al. 2014; Fang et al. 2020). Similar lack of genetic parallelism has been observed in Andean waterfowl for parallel adaptations in haemoglobin function (Natarajan et al. 2015) and in Neotropical cichlids for parallel phenotypic evolution of the visual system (Härer et al. 2018).

What is exceptional about the divergences of Arctic charr is their high degree of repeatability at the phenotypic level and the possibility to sample large numbers of 'evolutionary replicates', i.e. putatively independent divergence events (Elmer 2016). This allows for systematic assessments of the shared targets of natural selection in ecomorphs and differentiation between ecomorphs, which can be used to infer genes, functions, and genomic regions of potentially high relevance for the rapid diversifications in Arctic charr. Due to the large amounts of data in whole-genome data sets and the complexity of the Arctic charr genome (Christensen et al. 2018), signatures of population genetic metrics can be used as proxies for detecting regions affected by natural selection in the genome (Ellegren 2014). While applying single metrics to detect selection can be problematic, a

combination of metrics using changes in different properties of the genome can allow for more reliable detection of selection (Kern and Schrider 2018; Schneider et al. 2020). These inferences have been facilitated by methods based on genotype likelihoods, i.e. analyses that do not rely on the exact knowledge about alleles at each base in the genome (Nielsen et al. 2011; Korneliussen et al. 2014).

Here, we made use of the recent developments in the analysis of natural selection using low-coverage whole-genome resequencing in six lake populations of Arctic charr, including five that show parallel phenotypic divergence into benthic and pelagic ecomorphs and two of these with additional piscivorous or other putatively benthivorous ecomorphs (Alexander and Adams 2000; Jacobs et al. 2020). We explored the genome-wide diversity and divergence among these populations, including genome-wide diversification, and inferred the influence of ancestral introgression. To identify the shared genetic bases of ecomorph divergence in this widespread natural context, we screened for parallelism in the genomic regions and specific genes showing signals of being under natural selection, the types and differentiation of substitutions in selected genes, and the gene functions enriched in regions under selection using a range of metrics. We found five genes, four of them involved in nervous system function and one in oogenesis and embryonic development, to be under shared selection across several methods and all ecomorph pairs. These genes are part of large-scale genomic regions under selection and overlap with the genomic location of ecologically relevant traits. We found genomic regions under differentiation and natural selection between sympatric ecomorphs across replicates to be enriched for functions associated with gene expression and melanosome-related processes.

4.3 Materials & methods

4.3.1 DNA extraction & testing of library preparation methods

We extracted high molecular weight DNA from fin clips or muscle tissue using the Macherey-Nagel NucleoSpin Tissue kit, following the standard protocol for non-clinical samples. After checking the quality and fragmentation status of the DNA on an electrophoresis gel and measuring the DNA concentration on a Qubit fluorometer using the dsDNA BR Assay (Life Technologies), we used a modified version of a published Nextera DNA Library Prep (Illumina) protocol for low-coverage whole-genome sequencing (WGS) (Therkildsen and Palumbi 2017).

We used a test run using four individuals on the Illumina NextSeq500 platform at Glasgow Polyomics (University of Glasgow) to establish the efficacy and robustness of the library preparation protocol and to check library quality and mapping success. The Nextera trials showed inconsistent fragment sizes and low DNA yield. Pippin Prep (Sage Science) and other fragment size selection methods proved unsuccessful in obtaining sufficiently high DNA concentrations, which is why we then trialled a more promising library preparation method using the NEBNext Ultra II FS DNA Library Prep Kit for Illumina (New England BioLab), with which we obtained more consistent fragment sizes and higher amounts of DNA.

4.3.2 Library preparation & whole-genome sequencing

We used NEBNext Ultra II FS DNA Library Prep Kit for Illumina (New England BioLab) to generate libraries for individuals from lochs Dùghaill (13 benthic/profundal and 13 pelagic fish), Uaine (which has a shared demographic history with Loch Dùghaill, Jacobs et al. 2020), Awe (three ecomorphs: 9 benthic fish, 9 pelagic fish, 4 fish of recently discovered putatively benthic ecomorph), Tay (two ecomorphs: 5 benthic fish, 7 pelagic fish), Rannoch (three ecomorphs: 4 benthic fish, 6 pelagic fish, 6 piscivorous fish), and na Sealga (two ecomorphs: 6 benthic fish, 6 pelagic fish) (total N = 92) (Figure 4.1). The sampled individuals included males and females in equal proportions. All whole-genome libraries were sequenced on the Illumina HiSeqX platform at BGI, Hong Kong using 2x150 bp reads at a coverage of ~3X per individual (based on reference genome size) except lochs Dùghaill and Uaine, which were sequenced at a coverage of ~5X per individual.

4.3.3 Read filtering, mapping, & calculation of genotype likelihoods

The data set was quality-checked with FastQC (Andrews 2010), SAMtools (Li et al. 2009), and Picard (<http://broadinstitute.github.io/picard>). Reads were filtered and library adaptors were removed with Trimmomatic (Bolger et al. 2014) using an ILLUMINACLIP seed mismatch threshold of 2, a palindrome clip threshold of 30, a simple clip threshold of 10, and a minimum adaptor length of 7. Leading and trailing bases with a PHRED score below 20 were removed. A 4-bp sliding window approach was used to remove reads that contained windows with a PHRED score below 20. A minimum read length of 30 bp was specified in Trimmomatic. Remaining adaptor remnants identified using FastQC were removed using cutadapt with an error tolerance of 0 (Martin 2011). After filtering, the cleaned reads were then again checked using FastQC to ensure that median read quality

PHRED scores were above 30 and to confirm the absence of adaptors and sequencing artefacts. The filtered and cleaned reads were then mapped against the Arctic charr reference genome version GCF_002910315.2_ASM291031v2 (Christensen et al. 2018) using the mem algorithm of BWA (Li and Durbin 2010). After sorting of BAM files with SAMtools, deduplication using GATK4.0.8.1 (McKenna et al. 2010), sequence library creation using Picard, and insertion-deletion (INDEL) target creation using all individuals as well as INDEL realignment using GATK3.8.1, we used ANGSD (Korneliussen et al. 2014) for site and variant filtering and genotype likelihood estimation.

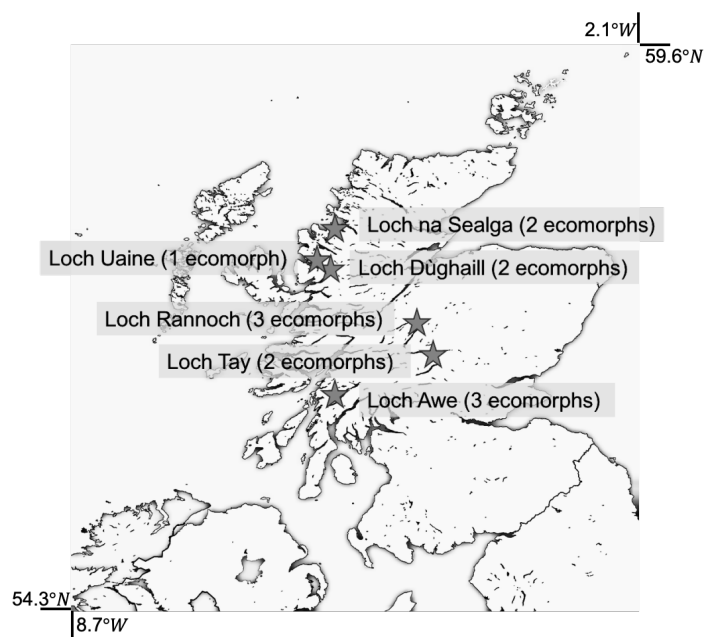


Figure 4.1 – Sampling locations of Arctic charr. The number of ecomorphs included per lake is shown in parentheses and the lake locations are marked with stars. Map: Creative Commons License (CC BY 2.5), www.demis.nl

Using all 92 individuals, we performed initial filtering of variant sites in ANGSD. To keep sites based on all populations in this filtering step, we specified the following criteria: 1) present in at least 61 of 92 individuals, 2) a minor allele frequency above 5%, 3) a PHRED quality score equal or higher than 20, 4) a minimum depth of 92 (equivalent to 1X depth of coverage per individual), 5) a maximum depth of 1840 (20 times the minimum depth), 6) a SNP variant *P* value less than 1E-6, 7) a mapping quality of at least 20, and 8) reads with a read flag above 255 (i.e. -remove_bads 1 option). After this initial global filtering, we performed filtering in ANGSD separately for each ecomorph within a population using the following criteria for keeping variant sites: 1) present in at least 50% of the individuals, 2) a minor allele frequency above 5%, 3) a PHRED quality score equal or higher than 20, 4) a minimum depth equal to the number of individuals in an ecomorph of a population, 5) a

maximum depth 20 times the minimum depth, 6) a mapping quality of at least 20, 7) reads with a read flag above 255, and 8) maximum of two alleles (i.e. biallelic). Genotype likelihoods were calculated based on these filtered sites separately for each ecomorph of a population.

4.3.4 Population structure & admixture analyses

We obtained genotype likelihood files in the BEAGLE format (Browning and Browning 2007) by creating BEAGLE files in ANGSD separately for each chromosome and then combining them into one file. Based on this combined BEAGLE file and all 92 sequenced individuals, we then estimated population structure using PCAngsd (Meisner and Albrechtsen 2018). Due to the drop in explained variance in the scree plot from PC5 onwards, we used the first four principal components (PCs) to describe the population structure. Using the first four PC scores, we then plotted the PCAngsd results using a custom R script in R v3.5.1 (R Core Team 2018).

To elucidate patterns of admixture and introgression among individuals within and across populations, we used NGSadmix (Skotte et al. 2013) on the combined BEAGLE genotype likelihood file. We tested K values (where K is the number of 'ancestral' populations) from 2 to 16. We then visualised population structure using custom R scripts.

4.3.5 ABBA-BABA tests for introgression

Due to significant admixture in lochs Rannoch and Tay visible in our PCAngsd and NGSadmix analyses, both within and between populations, we performed multi-population ABBA-BABA tests (also called four-taxon or four-population tests) in ANGSD (Green et al. 2010; Durand et al. 2011; Soraggi et al. 2018). We used the benthic Loch na Sealga ecomorph as outgroup, due to its considerable distance from any individuals of the Tay system. The sites used for the ABBA-BABA tests were the ones derived from the global ANGSD filtering step above. In ANGSD, we used the following filtering and analysis options for running the ABBA-BABA test on the deduplicated and realigned BAM files: 1) a minimum read depth of 4, 2) a maximum read depth of 100 (25 times the minimum read depth), 3) a minimum of 5 individuals, 4) a minimum read PHRED score of 20, 5) a minimum mapping quality of 20, 6) an analysis block size of 5,000,000 bp, and 7) using both transitions and transversions. The test was run with a total of six 'populations' (3 Loch Rannoch ecomorphs, 2 Loch Tay ecomorphs, and the benthic Loch na Sealga ecomorph as outgroup). However, the actual likelihood ABBA-BABA tests were performed iteratively in population quartets. The R script `estAvgError.R` (DSTAT) provided with the ANGSD

package was used to calculate significance of D-statistics (i.e. significantly higher ABBA than BABA patterns or vice versa).

4.3.6 Metrics of diversity & selection across the genome

Based on the genotype likelihoods, we calculated unfolded site-frequency spectra (SFS) (with the reference allele as ancestral, since it belongs to another lineage of *S. alpinus*) (Christensen et al. 2018). Based on the obtained SFS, we then used ANGSD to calculate F_{ST} (relative differentiation) (Wright 1949) between benthic and pelagic ecomorphs of each lake population and, separately for each ecomorph, *Tajima's D* (Tajima 1989), and π (nucleotide diversity) (Nei and Li 1979). Based on the minor allele frequency files obtained using ANGSD, we also calculated allele frequency difference (AFD) as an alternative measure of relative differentiation, following Berner 2019, and d_{XY} (also referred to as absolute divergence) (Takahata and Nei 1985; Fumagalli et al. 2014). AFD was calculated using custom R scripts. d_{XY} was calculated using the script calcDxy (J. Peñalba, unpublished), which was developed to alleviate the biased d_{XY} calculation in ANGSD that has been suggested by the ANGSD authors (2016; calcDxy.R, available at <https://github.com/mfumagalli/ngsPopGen/blob/master/scripts/calcDxy.R>). Using custom R scripts, we then performed sliding window analyses for each of the metrics with a window size of 50,000 bp and step size of 10,000 bp.

Additional to the metrics above, we performed a SNP-based PCAngsd differentiation scan (Meisner and Albrechtsen 2018) based on the principal component that separated the benthic and pelagic ecomorphs of all five populations. This genomic selection scan is based on posterior expectations of genotypes and uses an extended model of the FastPCA (Galinsky et al. 2016) algorithm. We converted the chi-squared-distributed selection statistics (with a degree of freedom of 1) in R and applied a false discovery rate (FDR) correction. SNPs with FDR q values below 0.01 were deemed significantly under selection along PC3.

Selection or differentiation outlier windows were defined using genome-wide quantiles (for F_{ST} and AFD: windows >95% quantile; for *Tajima's D*: windows <5% quantile). Based on the genome annotation provided with the published reference genome (Christensen et al. 2018), we then extracted genes that overlapped with outlier windows (hereafter called 'outlier genes') using custom R scripts. The obtained outlier genes formed the basis for the analyses of sharing of genes under selection below. Additionally, we recorded the linkage group information as well as the genomic start and end positions of outlier windows for subsequent functional enrichment analyses.

4.3.7 Shared outlier genes & windows

We determined the extent to which loci putatively under selection (i.e. outlier loci) were shared across populations using two approaches. First, we obtained genes that were selection outliers in two, three, four, or five ecomorph pairs (i.e. shared outlier genes that overlapped with genomic windows under selection). Second, we determined physical overlap of outlier windows in two, three, four, or five ecomorph pairs, regardless of the genes they contained (i.e. shared outlier regions or windows for subsequent functional enrichment analyses). Shared outlier genes across the five benthic-pelagic ecomorph pairs as well as across metrics were obtained using custom UNIX Bash scripts. Shared outlier windows were determined using custom R scripts using the R package IRanges v2.16.0 (Lawrence et al. 2013) to assess the physical overlap of genomic window positions. We established the best physical overlap size using 1) absolute start and end positions of genomic windows (or position of SNP for PCAngsd genomic selection outliers), 2) start positions of genomic windows minus 100,000 bp and end positions plus 100,000 bp, and 3) start positions of genomic windows minus 200,000 bp and end positions plus 200,000 bp, and then used approach 3) to obtain more large-scale genomic windows under shared selection and differentiation, which are potential hot spots of genomic divergence and diversification in Arctic charr and which were analysed for enrichment of gene functions. For the metrics F_{ST} , AFD, and local PCA, sharing was determined based on benthic-pelagic ecomorph pairs in each population. For *Tajima's D*, outlier gene and window sharing was determined separately for the benthic and pelagic ecomorphs rather than in benthic-pelagic ecomorph pairs. After assessing sharing of outlier genes and windows across populations/ecomorph pairs, we determined sharing across methods (F_{ST} , AFD, *Tajima's D*, and PCAngsd differentiation scan), both in terms of sharing of genes under selection and sharing of outlier regions (genomic outlier windows) that were used for functional enrichment analyses.

4.3.8 Functional enrichment

For gene ontology (GO) term enrichment analyses, we first extracted zebrafish gene symbols of all genes in shared genomic regions under selection (based on approach 3 in 4.3.7) from the Arctic charr reference genome annotation file. We converted part of the gene symbols for which only the locus number and gene name, but not the gene symbol, were present in the annotation file based on the MGI vertebrate homology database (<http://www.informatics.jax.org/homology.shtml>) and UniProt/SwissProt gene information. A GO term overrepresentation analysis on the gene symbols was performed in

WebGestalt v2019 (<http://www.webgestalt.org/>; last access: 16.05.2020) (Zhang et al. 2005; Wang et al. 2013, 2017; Liao et al. 2019) with a zebrafish GO term reference database. For this, we used biological process GO terms and the following analysis parameters: 1) a minimum number of genes per category of 3, 2) a maximum number of genes per category of 2000, 3) a Benjamini-Hochberg FDR correction (Benjamini and Hochberg 1995), 4) retaining the top 10 overrepresented GO terms, 5) a number of categories expected from set cover of 10, and 6) a number of visualised categories of 40. GO terms with FDR q values below 0.05 were deemed significantly overrepresented, while those with P values below 0.05 but FDR q values above 0.05 were considered to show a trend of overrepresentation. We used these two categories of overrepresentation to avoid being too conservative in our inference, since many of the GO terms are only represented by a small number of genes.

4.3.9 Overlap of outlier genes with salmonid QTLs

To determine if outlier genes overlapped with ecologically relevant quantitative trait loci (QTLs) of a recently published salmonid database (Jacobs et al. 2017) referenced to the Atlantic salmon genome version GCF_000233375.1_ICASG_v2 (Lien et al. 2016), we extracted the Arctic charr reference genome sequence of the top shared outlier genes that were shared across all ecomorph pairs and at least two of the selection outlier methods. Then we used NCBI megablast (Altschul et al. 1990; Morgulis et al. 2008; Chen et al. 2015) against all Atlantic salmon (*S. salar*) (taxid:8030) gene sequences available on NCBI GenBank to obtain orthologous genes and gene positions for salmon. Due to mapping ambiguity in the megablast (blastn) results, we additionally used blastp (Altschul et al. 1990; Gish and States 1993) to obtain the orthologous proteins and gene positions of the top blastp hit (highest 'Max Score') for Atlantic salmon. After obtaining the Atlantic salmon-referenced outlier gene regions, we then determined the overlap of these regions with the 95% confidence intervals of all QTLs in the database (Jacobs et al. 2017) and recorded the type of QTL, the phenotypic trait, and the species the QTL was described in.

4.3.10 Types of substitutions in outlier genes

To classify the types of substitutions in outlier genes and identify in which gene regions the substitutions occurred, we used SNP functional effect analysis software SnpEff (Cingolani et al. 2012). We first created a SnpEff database for Arctic charr based on the GCF_002910315.2 gff3 annotation version 101 file and the GCF_002910315.2_ASM291031v2 reference genome sequence in fasta format. Next, we

transformed the BEAGLE file derived from all populations in ANGSD using custom Bash scripts to a vcf v4.4 file. SNPs were then annotated and SNP effects calculated based on this vcf file. We then recorded the exact locations and types of substitutions in the top five shared outlier genes (*grid1* on chromosomes 11 and 21, *diaph2*, *mdga2*, and *ptprn2*) and calculated the maximum AFD values of these substitutions between the benthic and pelagic ecomorphs averaged across populations.

4.3.11 Phylogenetic relationships of *grid* genes in Arctic charr

We examined the two versions of the gene *grid1* that were found to be under shared selection and differentiation on two separate chromosomes (NC_036848.1 = chromosome 11 and NC_036858.1 = chromosome 21). We downloaded the complete DNA sequences of the coding regions of all the available versions of *grid1* on GenBank as well as the sequences of *grid1a*, *grid1b*, and *grid2* of zebrafish and *GRID1* and *GRID2* of humans (Table S4.1). We used the ClustalW algorithm (Higgins and Sharp 1988; Larkin et al. 2007) implemented in MEGA X (Kumar et al. 1993, 2018; Stecher et al. 2020) using default options to align the sequences. Then we used the maximum likelihood phylogeny construction method in MEGA X to determine the relationships among the different versions of *grid*. We used 1000 bootstrap replicates to assess node support, a general time reversible (GTR) substitution model, uniform rates among sites, complete deletion of missing sites, nearest-neighbour interchange (NNI), automatic initial tree construction, and no branch swap filters.

4.3.12 Principal component analysis on metrics & their correlations

To obtain the similarities of selection and diversity metric signals across populations and ecomorph pairs, and to assess the relationship between pairs of metrics, we applied a PCA-based approach following Stankowski et al. (2019). First, we used the IRanges function to match all metric values in 50,000 bp sliding windows with a step size of 10,000 in one data frame using a custom R script. For this, we also performed a sliding window analysis of selection statistic values with the PCAngsd differentiation output. Next, PCA analysis using the prcomp function in R was applied to the z-transformed sliding window values of each metric across the five populations. After examining the variance explained by the different principal components, we focussed on PC1 for the downstream analyses and extracted the population loadings on the principal components to determine the relative contribution of each population to the PC scores. Using these PC1 scores as proxies for the metric values in all populations, we then calculated correlation coefficients between pairs

of metric PCs and performed correlation significance tests in R with the `cor.test` function. Additionally, we created a correlation and scatter plot matrix with the `pairs.panels` function in the R package `psych` v2.0.7 (Revelle and Revelle 2015; Revelle 2020).

4.4 Results

4.4.1 Mapping success and variant statistics

The average coverage of fish from lochs Dùghaill and Uaine was 4.95X and 4.32X, respectively (Table S4.2). Regions of the genome with at least 1X coverage had an average coverage of 10.26X (Dùghaill) and 8.77X (Uaine). For the other lakes, the average coverage was 2.54 (Awe), 2.47 (na Sealga), 3.67 (Rannoch), and 2.46 (Tay) (Table S4.2). Genomic regions with at least 1X coverage (covered regions) had an average coverage of around 6.27 (Awe), 5.92 (na Sealga), 9.04 (Rannoch), and 5.84 (Tay). Using ANGSD, we obtained a total of 15,935,467 SNP positions with genotype likelihoods for all 92 individuals of all lochs combined. A total of 1,952,881 SNP positions had no missing data across all individuals, while the former number included SNP positions with missing data.

4.4.2 Population structure & admixture

When assessing the population structure of all individuals (Figure 4.2), PC3 (explaining 14.2% of variance in allele frequencies) separated all the populations clearly and generally separated the ecomorphs within populations. PC1 (explaining 27.2% of variance) mainly separated Dùghaill and Uaine and to some degree Awe from the other three populations. PC2 (explaining 19.3% of variance) mainly separated na Sealga and to some degree Awe from the remaining populations. PC4 (explaining 9.1% of variance) clearly separated both Loch Dùghaill ecomorphs from Loch Uaine. Together, PCs 1 to 4 separated both the populations and the ecomorphs within populations, despite some limited overlap of ecomorphs within populations (Figure 4.2).

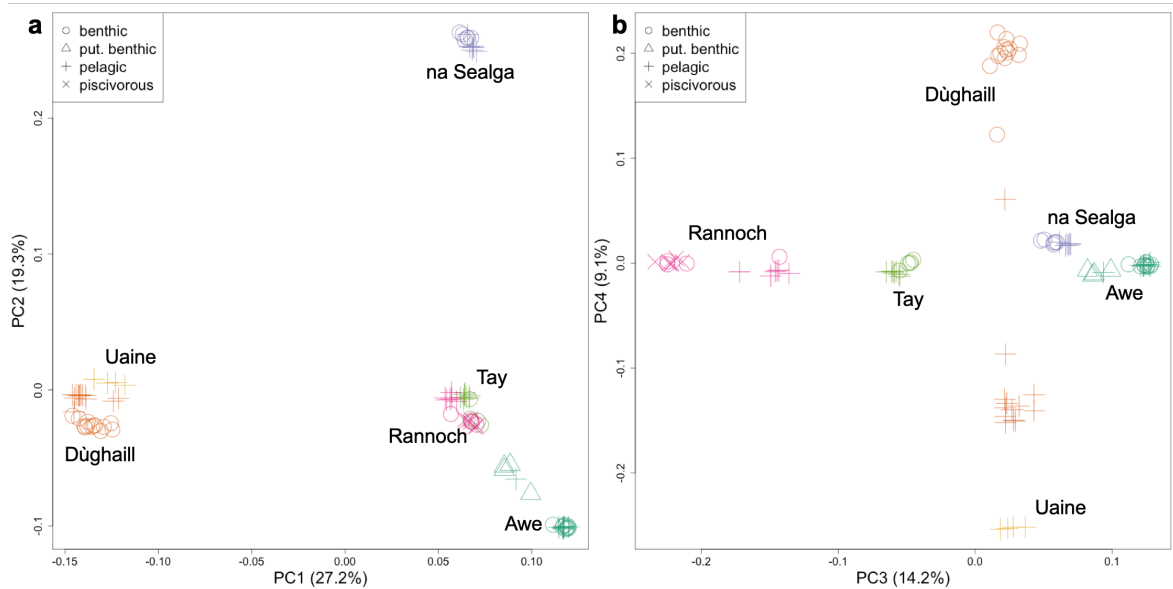


Figure 4.2 – PCAngsd population structure plots. a) Plot of principal component (PC) 2 against PC1. b) Plot of PC4 against PC3. The variance explained by each PC is shown in parentheses. benthic ecomorph = circles, putatively benthic ecomorph = triangles, pelagic ecomorph = plus signs, piscivorous ecomorph = crosses

We observed higher likelihood in NGSadmixture analyses with higher K values compared to lower K values (Table S4.3). The number of ancestral populations that most closely matched the population-ecomorph combinations in this study and that had higher likelihood values than lower K s was seven ($K = 7$; $L = -857,170,124$), as inferred from NGSadmixture (Figure 4.3). We observed the following clusters: 1) The benthic and pelagic ecomorphs of Loch Awe, 2) the recently discovered putatively benthivorous ecomorph of Loch Awe, 3) the pelagic ecomorphs of lochs Dùghaill and Uaine, 4) the benthic/profundal ecomorph of Loch Dùghaill, 5) both ecomorphs of Loch na Sealga, 6) all three ecomorphs of Loch Rannoch, and 7) both ecomorphs of Loch Tay. Higher values of K led to the same clustering by population and similar clustering by ecomorph as $K = 7$ but higher substructuring within the ecomorphs of lakes, for instance for $K = 16$, which had the highest log-likelihood among all K s ($L = -799,178,303$) (Figure S4.1).

However, judging from the NGSadmixture admixture patterns, irrespective of the chosen K value, there is evidence for hybridisation, introgression or, alternatively, pronounced retained ancestral variation in benthic-pelagic ecomorph pairs of three of the populations (Dùghaill, Rannoch, and Tay). The ABBA-BABA tests in the Tay system (Rannoch and Tay) showed that the benthic ecomorphs of the two lakes share substantial genetic variation, either through introgression or common ancestry, while the pelagic ecomorphs do not (according to significance and frequency of ABBA-BABA patterns in

four-taxon tests; Table S4.4). Overall genome-wide differentiation, measured as allele frequency difference (AFD), was highest between the benthic and pelagic ecomorphs of Loch Tay (mean AFD = 0.1390), in the benthic-pelagic ecomorph pair in Loch Rannoch (mean AFD = 0.1339), and between the benthic and pelagic ecomorphs in Loch Dùghaill (mean AFD = 0.1284). The lowest levels of overall differentiation were observed in Loch na Sealga (mean AFD = 0.1039) and between benthic and pelagic ecomorphs in Loch Awe (mean AFD = 0.0975).

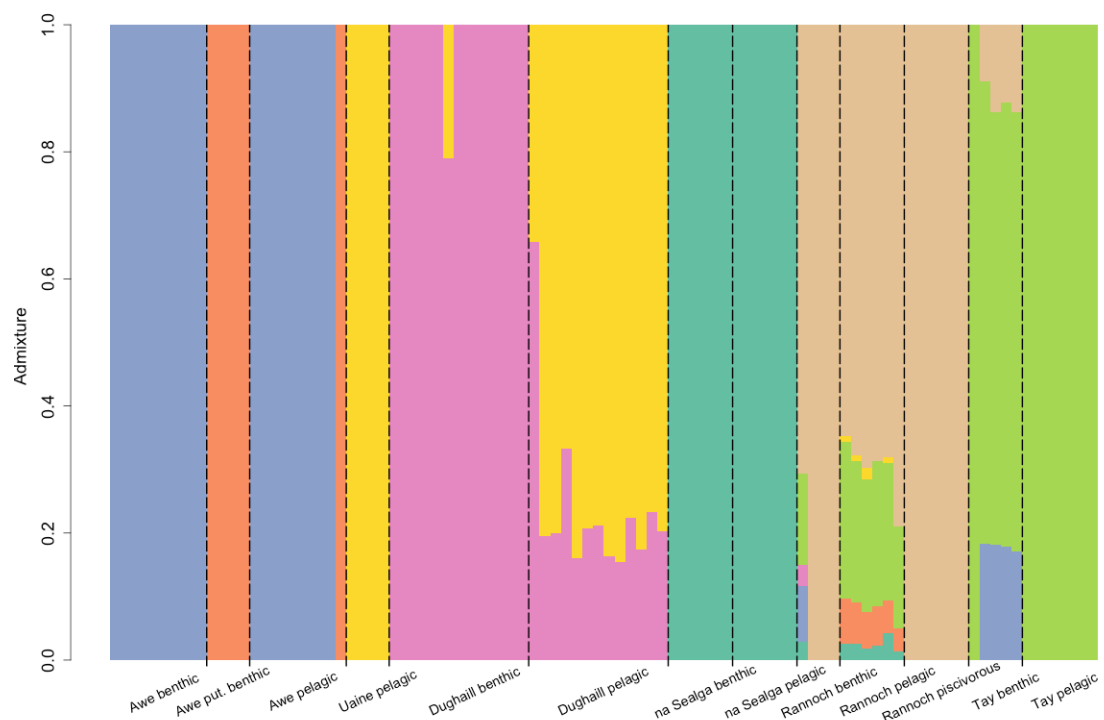


Figure 4.3 – Individual genome-wide admixture patterns in Arctic charr across six lakes and up to three sympatric ecomorphs. The y-axis represents the proportion of the genome that clusters with a certain assumed ancestral population (K). The plot for an assumed ancestral number of populations of $K = 7$ is shown. Awe put. benthic = putatively benthic ecomorph of Loch Awe

4.4.3 Shared selection & differentiation outliers in benthic-pelagic ecomorph pairs

We found several outlier genes that were shared across all five benthic-pelagic ecomorph pairs in two or more of the selection and differentiation metrics. The AFD and PCAngsd differentiation scan methods detected two outlier genes across all benthic-pelagic ecomorph pairs: *ptprn2* and *grid1* (on chromosome NC_036858.1 = chromosome 21) (Table 4.1). For F_{ST} , PCAngsd, and *Tajima's D* in the benthic ecomorphs, we also found

grid1 as one of the shared outlier genes, albeit a different variant of the gene on a different chromosome (chromosome NC_036848.1 = chromosome 11), as well as *diaph2*. For F_{ST} , PCAngsd, and *Tajima's D* in the pelagic ecomorphs, only *grid1* on chromosome 11 remained as a shared outlier gene. Finally, we found three genes to be outliers in both F_{ST} and PCAngsd differentiation analyses across all five benthic-pelagic ecomorph pairs: *grid1* on chromosome 11, *diaph2*, and *mdga2* (Table 1). Thus, *grid1* was an outlier gene in all differentiation and selection analyses across all five benthic-pelagic ecomorph pairs, albeit on a different chromosome in the AFD analysis. However, both of these variants of the gene *grid1* were outliers in the PCAngsd differentiation analysis (FDR q value < 0.01).

Table 4.1 – Shared outlier genes across all five benthic-pelagic ecomorph pairs of Arctic charr. AFD and F_{ST} outliers are above the top 5% quantile. *Tajima's D* outliers are below the bottom 5% quantile. PCAngsd differentiation outliers have FDR q values below 0.01. *Tajima's D* outliers are ecomorph specific. Chromos. = Chromosome; AFD = allele frequency difference; outl. = outlier; TajD = *Tajima's D*; benth. = benthic; pelag. = pelagic; chr = chromosome

Gene symbol	Gene name	Chromos.	Position [Mb]	AFD outl.	F_{ST} outl.	TajD benth. outl.	TajD pelag. outl.	PCAngsd outl.
<i>diaph2</i>	protein diaphanous homolog 2	NC_036848.1 = chr 11	24.7		yes	yes		yes
<i>grid1</i>	glutamate receptor ionotropic, delta-1	NC_036848.1 = chr 11	31.8		yes	yes	yes	yes
<i>grid1</i>	glutamate receptor ionotropic, delta-1	NC_036858.1 = chr 21	6.3	yes				yes
<i>mdga2</i>	MAM domain-containing glycosylphosphatidylinositol anchor protein 2	NC_036868.1 = chr 31	25.4		yes			yes
<i>ptprn2</i>	receptor-type tyrosine-protein phosphatase N2	NC_036854.1 = chr 17	48.9	yes				yes

4.4.4 Phylogenetic relationships of *grid* genes

A phylogenetic analysis of grid genes revealed that there are two copies of the grid gene in Arctic charr, which cluster with *grid1a* of zebrafish, and two that cluster with *grid1b* of zebrafish (Figure S4.2). One of the *grid1a* genes was the shared outlier gene on chromosome 21 (NC_036858.1), the other gene being a copy of *grid1a* on chromosome 28 (NC_036865.1) that was not a shared selection outlier in our analyses. Similarly, one of the *grid1b* genes was the shared outlier gene on chromosome 11 (NC_036848.1), the other gene being a copy of *grid1b* on chromosome 21 (NC_036858.1) that was not a shared selection outlier. The human version of *grid1* formed a sister clade with the *grid1* genes of charr and zebrafish (Figure S4.2). The *grid2* genes of zebrafish and human were sister to all the *grid1* genes included in the phylogeny.

4.4.5 Nucleotide substitutions around outlier genes

There were no nucleotide substitutions in the coding regions of the top five shared outlier genes *grid1* (on chromosomes 11 and 21), *diaph2*, *mdga2*, and *ptprn2*, as inferred from the SnpEff analysis. However, we found a substantial number of substitutions in the upstream and downstream regions (± 5 kb) around the genes, none of which were fixed. There were 21 upstream and 74 downstream gene variants around *grid1* on chromosome 11 (NC_036848.1), with maximum AFD values between the benthic and pelagic ecomorphs averaged across populations of 0.13 and 0.33, respectively. For *grid1* on chromosome 21 (NC_036858.1), we found 16 up- and 8 downstream gene variants (max. AFD = 0.20 and 0.12, respectively). Around the gene *diaph2*, we found a total of 31 upstream gene variants, many of which were in close proximity, and 11 downstream gene variants (max. AFD = 0.37 and 0.19, respectively), and for *mdga2* we found 24 up- and 21 downstream gene variants (max. AFD = 0.23 and 0.33, respectively). Lastly, we found 12 up- and 2 downstream gene variants around *ptprn2* (max. AFD = 0.26 and 0.20, respectively).

4.4.6 Relationships between population genetic metrics

We examined relationships between metrics across populations and between pairs of metrics using a combination of principal component analysis and correlation tests. The first principal components of metrics across populations (AFD, F_{ST} , Tajima's D , d_{XY} , π) or benthic-pelagic ecomorph pairs explained most of the variation, indicating strong sharing of metric signals, with relatively equal distribution of principal component loadings across populations (Table S4.5). We found a strong correlation between the PC1s of differentiation metrics F_{ST} and AFD ($r = 0.87$, $P = 2.2E-16$; Figure S4.3; Table S4.6). In

contrast, the correlation between PC1s of F_{ST} and absolute divergence (d_{XY}) was low and positive ($r = 0.14$, $P = 2.2E-16$), while the correlation between the PC1s of AFD and d_{XY} was low but negative $r = -0.20$, $P = 2.2E-16$; Figure S4.3; Table S4.6). The correlation between the PC1s of F_{ST} and nucleotide diversity (π) in the benthic and pelagic ecomorphs was positive and similarly low ($r = 0.20$ and 0.19 , $P = 2.2E-16$ and $2.2E-16$, respectively), while the correlation between the PC1s of AFD and π was negative but close to zero ($r = -0.08$ and -0.08 , $P = 2.2E-16$ and $2.2E-16$, respectively).

4.4.7 Functional overrepresentation in genomic regions of shared selection

A total of 130 shared outlier regions (50K bp windows + 200K bp up- and – 200K bp downstream each, spanning a total sequence length of 58.5 Mb; approach 3 in 4.3.7) across AFD, PCAngsd, and *Tajima's D* in benthic ecomorphs across all five main populations were found on 15 of 39 chromosomes (Figure 4.4). These outlier regions contain 85 annotated genes. While we did not find significant functional enrichment after correction for multiple testing (FDR q value = 1), we found strong trends of overrepresentation ($P < 0.001$) for functions (GO terms) involved in 1) the regulation of gene expression and 2) melanosome metabolism and transport (top 10 overrepresented functions: Table 4.2). Similarly, for AFD, PCAngsd, and *Tajima's D* in pelagic ecomorphs, we obtained a total number of 113 shared outlier regions (spanning a total sequence length of 50.85 Mb; approach 3 in 4.3.7) on 18 of 39 chromosomes across all five main populations (Figure S4.4). Again, we did not find significant functional enrichment after correction for multiple testing (lowest FDR q value = 0.7666). However, several GO terms related to gene expression showed strong trends of overrepresentation ($P < 0.001$; top 10 overrepresented functions: Table 4.2).

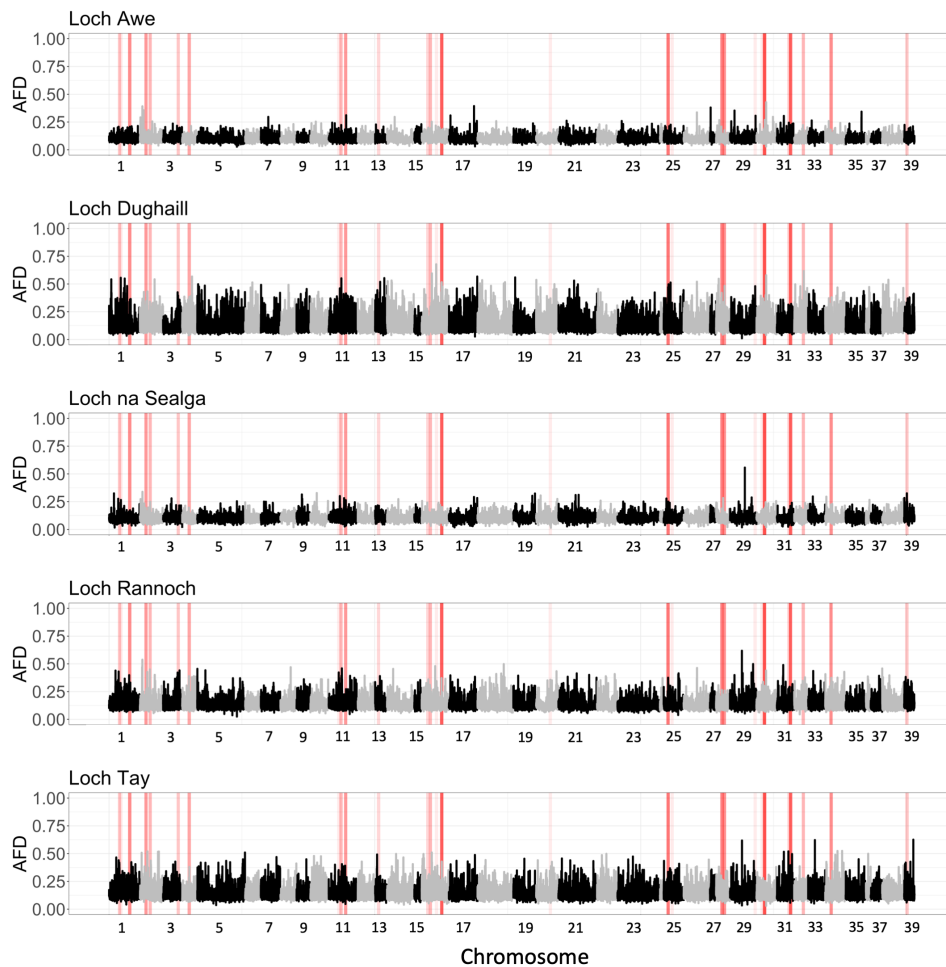


Figure 4.4 – Allele frequency difference (AFD) between benthic and pelagic ecomorphs. The landscape of AFD is shown separately for all five benthic-pelagic ecomorph pairs. The grey or black shadings indicate the 39 assembled chromosomes of Arctic charr. Chromosome ID numbers are shown below the AFD distributions. The red bars indicate shared outlier regions across AFD (top 5% quantile) and PCAngsd (FDR q value < 0.01) analyses across all five benthic-pelagic ecomorph pairs and *Tajima's D* (bottom 5% quantile) in the benthic ecomorphs of the five main populations (approach 3 in 4.3.7). Unassembled portions of the genome, i.e. those with uncertain genomic location, were not included in these plots.

4.4.8 Overlap of outlier genes with ecologically relevant QTLs

All of the five outlier genes shared across all five benthic-pelagic ecomorph pairs overlapped with known ecologically relevant QTLs in salmonids (Table S4.7). The *grid1* gene on chromosome 11 (NC_036848.1) overlapped with two QTLs of *S. alpinus* involved in body length and growth rate, respectively, as well as with three QTLs of other salmonid genera involved in sexual maturation and embryonic development. The *grid1* gene on chromosome 21 (NC_036848.1) overlapped with two QTLs of *S. alpinus* involved in head

shape and condition factor, respectively, as well as with one QTL of *Oncorhynchus mykiss* involved in body weight. The *diaph2* outlier gene overlapped with a body weight QTL of *S. alpinus*. The *mdga2* gene overlapped with a body length QTL of *Oncorhynchus tshawytscha*. Lastly, the *ptprn2* outlier gene overlapped with a body shape QTL of *Oncorhynchus mykiss*.

Table 4.2 – Top 10 overrepresented gene ontology (GO) terms in shared selection outlier regions (approach 3 in 4.3.7). The most highly overrepresented GO terms of outlier regions shared across AFD and PCAngsd differentiation analyses in all five benthic-pelagic ecomorph pairs and *Tajima's D* analyses in the benthic ecomorphs (top half) and pelagic ecomorphs (bottom half) across populations. Description (biological process) = name of biological process GO term, Observed = number of times GO term occurred in outlier regions, Expected = number of times a GO term is expected to appear in genes of outlier regions, Ratio = ratio of observed to expected number of times a GO term appears in genes of outlier regions, *P* value = uncorrected *P* value, FDR = false discovery rate (Benjamini-Hochberg correction) *q* value

GO term	Description (biological process)	Observed	Expected	Ratio	<i>P</i> value	FDR
Benthic ecomorphs						
GO:0061158	3'-UTR-mediated mRNA destabilisation	8	0.0299	66.789	3.8E-04	1
GO:0050779	RNA destabilisation	12	0.0449	44.526	8.8E-04	1
GO:0061157	mRNA destabilisation	12	0.0449	44.526	8.8E-04	1
GO:0061014	positive regulation of mRNA catabolic process	18	0.0674	29.684	2.0E-03	1
GO:0032401	establishment of melanosome localisation	18	0.0674	29.684	2.0E-03	1
GO:0032402	melanosome transport	18	0.0674	29.684	2.0E-03	1
GO:0051904	pigment granule transport	18	0.0674	29.684	2.0E-03	1
GO:0032400	melanosome localisation	19	0.0711	28.122	2.2E-03	1
GO:0051905	establishment of pigment granule localisation	20	0.0749	26.716	2.5E-03	1
GO:0043412	macromolecule modification	1725	6.4569	2.1682	2.6E-03	1
Pelagic ecomorphs						
GO:0061158	3'-UTR-mediated mRNA destabilisation	8	0.0213	94	1.9E-04	0.76 660
GO:0050779	RNA destabilisation	12	0.0319	62.667	4.4E-04	0.76 660
GO:0061157	mRNA destabilisation	12	0.0319	62.667	4.4E-04	0.76 660
GO:0061014	positive regulation of mRNA catabolic process	18	0.0479	41.778	1.0E-03	1

GO:1903311	regulation of mRNA metabolic process	78	0.2074	14.462	1.1E-03	1
GO:0006518	peptide metabolic process	322	0.8564	5.8385	1.4E-03	1
GO:0043488	regulation of mRNA stability	23	0.0612	32.696	1.7E-03	1
GO:0043487	regulation of RNA stability	26	0.0691	28.923	2.1E-03	1
GO:1903313	positive regulation of mRNA metabolic process	27	0.0718	27.852	2.3E-03	1
GO:0061013	regulation of mRNA catabolic process	30	0.0798	25.067	2.8E-03	1

4.5 Discussion

We found that selection and ecomorph differentiation are parallel at only a small proportion of the genome across the five replicate benthic-pelagic divergences but that this is enriched for relevant genes to a biologically important extent. Five protein-coding genes were highly shared across methods to detect selection and across all ecomorph pairs. Two of these genes were independent copies of the *grid1* gene on different chromosomes, while the other three were *diaph2*, *mdga2*, and *ptprn2*. Moreover, all of the five shared outlier genes overlapped with one or more known ecologically relevant QTLs (Jacobs et al. 2017). At the level of larger genomic regions, sharing of selection targets was higher than at the gene level. The most highly overrepresented gene functions in these regions mainly included processes involved in the regulation of gene expression, while in benthic charr they also included functions with a role in melanosome metabolism and transport.

4.5.1 Relationships among divergences & origin of selected genetic variation

Based on our population structure, admixture, and introgression analyses, the divergence histories of these Arctic charr ecomorphs are complex and vary considerably among lakes. Both Awe and na Sealga are likely cases of sympatric speciation (i.e. ecomorphs likely differentiated independently from ecomorphs of other lakes), evidenced by their low genetic differentiation between ecomorphs, lack of admixture across lakes, and in agreement with historical demographic reconstructions (Jacobs et al. 2020). In the case of Loch Dùghaill, the high relatedness between the pelagic ecomorphs of lochs Dùghaill and Uaine suggest an allopatric phase followed by secondary contact of the two ecomorphs of Loch Dùghaill, which is in agreement with evidence from RAD-seq data (Jacobs et al. 2020). The origin and original state and ecology of the ecomorphs of lochs Dùghaill and Uaine is unclear and warrants further research. For lochs Rannoch and Tay, both of which are part of the Tay basin, we found evidence for introgression or shared ancestry between the two lakes, particularly for the benthic ecomorphs, in line with previous studies on this

lake system that showed evidence of gene flow between the two lakes (Verspoor et al. 2010), which is potentially facilitated by connected river systems between the two lakes. In contrast to lochs Awe and na Sealga, the divergence and differentiation events of ecomorphs in lochs Rannoch and Tay and similarly for Loch Dùghaill cannot readily be considered independent, while divergence in sympatry might still play an important role in the build-up of differentiation. A role of introgression in facilitating ecomorph divergence seems particularly likely in the case of lochs Rannoch and Tay, for which we found evidence for across-lake introgression. Whether the observed introgression can be considered adaptive, as has been found in e.g. cichlids (Malinsky et al. 2018) and tropical butterflies (Enciso-Romero et al. 2017), cannot currently be determined. Similar introgression is also possible between lochs Dùghaill and Uaine (discussed in Chapter 5). Additional studies and sampling of further populations of Arctic charr at a broader geographic scale are needed to gain a deeper understanding about the phylogeography, relationships, and differentiation of ecomorphs across Scotland and beyond. Determining the origin of extant ecomorphs of Arctic charr and the amount of retained ancestral genetic variation would allow conclusions about the independence or non-independence of the divergences of ecomorphs of Arctic charr included here.

Our findings highlight a strong common element of adaptation at the genomic level, evidenced by the sharing of differentiation and selection outlier genes in benthic-pelagic ecomorph pairs, the overlap of outliers with ecologically relevant QTLs, and the evidence of frequent positive selection on the recovered outlier genes in other systems. This parallel evolution at the genomic level has occurred despite the often complex and variable demographic backgrounds in Arctic charr populations and ecomorphs. However, despite higher sharing of larger genomic regions under selection (approach 3 in 4.3.7) and strong trends of functional overrepresentation in these regions, sharing at the gene level in this system is low, implying an important role of polygenic adaptation leading to phenotypic parallelism (Brawand et al. 2014; Laurentino et al. 2020). This result is also in line with high predictability and parallelism at the phenotypic level, but low predictability at the specific genetic level (Natarajan et al. 2015; Bolnick et al. 2018; Jacobs et al. 2020).

Given the recent age of the studied divergences, ranging from 3.7 to 50.1 Kya (Jacobs et al. 2020), selection on *de novo* genetic variants is unlikely to be the main driver of adaptive genomic differentiation. Combined with the evidence of strong correlation of both *Tajima's D* and π signals between the benthic and pelagic ecomorphs, our findings suggest an important role of shared ancestral variation in divergence along the benthic-pelagic axis, despite the observed low sharing at the gene level. Moreover, the lack of correlation between differentiation and nucleotide diversity suggests that differentiation

between ecomorphs is not affected by low diversity to the same extent as in other studies, which makes it more likely that differentiation is indeed shaped by a response to natural selection (Cruickshank and Hahn 2014; Ravinet et al. 2017; Wolf and Ellegren 2017). Lastly, the low correlation between relative differentiation (AFD and F_{ST}) and absolute divergence (d_{XY}) indicates a recent origin of differentiation across the genome, despite the likely ancestral origin of the selected genetic variants (Cruickshank and Hahn 2014; Ravinet et al. 2017; Wolf and Ellegren 2017). Therefore, similar to the case of recurrent marine vs freshwater divergence in three-spined stickleback (Jones et al. 2012; Roesti et al. 2014; Nelson and Cresko 2018; Kirch et al. 2020), positive selection on standing or ancestral genetic variation is likely to be crucial for genomic adaptation for some but not all of the diverged traits in the postglacial diversifications of Arctic charr. Comparable evidence comes from the divergence into ecomorphs in whitefish (*Coregonus*), where shared selection on standing genetic variation has been found in six sympatric ecomorph pairs in Europe and North America (Rougeux et al. 2019). Such reuse of existing genetic variants underlines the potential of standing genetic variation for parallel ecological divergence.

4.5.2 Sharing of selection outliers reveals shared ecological challenges

Our analyses indicated a low frequency of but nevertheless potentially biologically significant sharing of genes under selection, despite our criteria for 'parallelism' being very stringent; in our study we only considered shared outlier genes to be those found across all ecomorph pairs and at least two of the four analytical methods used to infer selection. In contrast, some other studies of parallel ecological adaptation consider even as low as a single pair from several being parallel (e.g. Stuart et al. 2017; Rennison et al. 2019). Therefore, we have high confidence that our top genes under selection are very likely to play key and pervasive roles in parallel ecological adaptation across independent replicates.

The shared outlier genes we identified are likely of high importance to the process of benthic and pelagic adaptation and point to the shared ecological challenges across populations that result in shared selection pressures. It is striking that four of the five shared outlier genes play a role in nervous system processes, and the roles these genes could play in ecological adaptation is evident from other research. *Grid1* in particular has been previously repeatedly reported in ecological divergences of fishes. In three-spined stickleback (*Gasterosteus aculeatus*), *grid1* was identified as a candidate gene for freshwater vs marine adaptation (Ferchaud et al. 2014) and in Atlantic salmon (*S. salar*) a

SNP in *grid1* was correlated with population decline (Lehnert et al. 2019). Moreover, a gene of the same family of glutamate receptors as *grid1*, *grm4*-like, was found as an outlier in a whole-genome selection scan comparing migratory and resident populations of brown trout (*Salmo trutta*) (Lemopoulos et al. 2018). We speculate that in Arctic charr, *grid1* could be under selection as a consequence of foraging behaviour and locomotion differences between benthic and pelagic charr, which likely pose differing selection pressures on synaptic plasticity. However, selection on *grid1* could also arise due to temperature or water pressure adaptation as opposed to behavioural adaptation alone, since benthic ecomorphs of Arctic charr often inhabit deeper parts of the lake for most of the year (Gardner et al. 1988; Adams et al. 1998; Fraser et al. 1998; Adams et al. 2007; Hooker et al. 2016).

Two other genes with nervous system functions under shared selection and differentiation across ecomorph pairs were *mgda2* and *ptprn2*. *Mdga2* is a gene with a putative role in neural cell migration in medaka (*Oryzias latipes*) (Sano et al. 2009). Furthermore, it was associated with growth performance in a QTL study in common carp (*Cyprinus carpio*) (Su et al. 2018), which could imply a putative role in foraging behaviour and the regulation of food intake in fish. This relationship between neural regulation and growth has been suggested due to the role of brain regulation in energy balance (Rui 2013; Rathjen et al. 2017; Su et al. 2018), which would also explain the overlap of *grid1* and *mdga2* with growth-related QTLs in our study (*grid1*: overlap with condition factor and body growth QTLs; *mdga2*: overlap with body length QTL). The fourth shared outlier gene involved in neural processes, *ptprn2*, plays a role in insulin secretion (Cui et al. 1996; Bence 2013; Carmona et al. 2014; Bian et al. 2017) and may also have a role in body growth (Lee 2019). Moreover, *ptprn2* was found to be both differentially expressed and to have undergone a hard selective sweep in *Cyprinodon* pupfishes in a sympatric radiation (McGirr and Martin 2020). In Atlantic salmon, a structural variant linked to *ptprn2* has been found to be associated with domestication (Bertolotti et al. 2020). In Arctic charr, a gene of the same family, *ptprj*, was found to contain shared ecomorph-associated SNPs (Jacobs et al. 2020).

The fifth of the shared outlier genes across ecomorph pairs, *diaph2*, has been implicated in oogenesis in humans (Bione et al. 1998) and, more generally, in actin cytoskeleton organisation (Gasman et al. 2003) as well as in early embryonic development in zebrafish (*Danio rerio*) (Lepage and Bruce 2010). It has been found to be under selection linked to local adaptation in the red fox (*Vulpes vulpes*) (Roberts 2019) and European rabbits (*Oryctolagus cuniculus*) (Carneiro et al. 2013). In grass carp (*Ctenopharyngodon idellus*) cultured at different stocking densities, *diaph2* has been found

to be differentially expressed in muscle tissue (Zhao et al. 2019). The specific role of *diaph2* and the other shared outlier genes in Arctic charr adaptation remains to be tested.

While there were no changes in the coding sequence of shared outlier genes, we found a high number of variants in the up- and downstream regions. These variants in the proximity of genes, especially in the upstream region, potentially regulate gene expression as *cis*-acting manipulators of transcription (Pai and Gilad 2014; Signor and Nuzhdin 2018), such as has been found for speciation- or divergence-driving traits in cichlids (Santos et al. 2014; York et al. 2018) as well as for salinity tolerance and ecomorph divergence in Arctic charr (Norman et al. 2014; Jacobs and Elmer 2020). In sticklebacks, whole-genome and RNA-seq data suggested that regulatory changes and particularly *cis*-regulatory variants are the main drivers of parallel ecomorph divergence (Jones et al. 2012; Verta and Jones 2019). Given that we did not find complete differentiation of any of the variants around the shared outlier genes, gene expression in these genes is likely a product of multiple variants, each with a putatively small effect, which has also been observed in lake-stream divergences in freshwater-inhabiting three-spined stickleback (Laurentino et al. 2020) and in the marine snail *Littorina saxatilis* (Westram et al. 2014, 2016, 2018; Johannesson and Butlin 2017; Morales et al. 2019).

4.5.3 Hot spots of differentiation & sharing of selected gene functions

In many cases of sympatric divergence, the genomic landscape of differentiation is not the result of single, isolated spikes of differentiation but rather contains islands of differentiation with often large numbers of genes in physical proximity along the genome, while this also depends on the time since the onset of divergence (Renaut et al. 2013; Malinsky et al. 2015; Larson et al. 2017; Wolf and Ellegren 2017; Papadopulos et al. 2019). While shared outliers in the same genomic location of the genome across populations are more likely to pinpoint genes of high importance, larger-scale genomic divergence hot spots can highlight genomic regions putatively associated with evolutionary adaptation (Roesti et al. 2014). Since these hot spots of divergence often show enrichment of certain gene functions (Laurent et al. 2017; Lawson and Petren 2017; Papadopulos et al. 2019), we analysed the shared genomic regions under selection in our data set for overrepresented GO terms that might offer a clue as to what biological functions are important for adaptation to benthic and pelagic environments across populations of Arctic charr.

In both the benthic and pelagic ecomorphs, functions related to the regulation of gene expression dominated among the most highly enriched GO terms. The importance of

gene regulation evolution for the speciation process has been found in various instances of ecology-driven divergence (Pavey et al. 2010; Mack et al. 2016). It has been suggested that positive selection on regulatory genes and compensatory evolution of regulatory divergence can aid in the build-up of reproductive isolation, since the evolution and divergence of regulatory genes makes Dobzhansky-Muller incompatibilities more likely (Mack et al. 2016). This is because detrimental epistatic interactions and misexpression in hybrids are common causes of hybrid sterility and postzygotic reproductive isolation (Mack et al. 2016; Morgan et al. 2020), which makes regulatory genes key drivers of sympatric speciation (Pavey et al. 2010; Mack and Nachman 2017).

In the benthic ecomorph of Arctic charr, we also found functions related to melanosome transport and metabolism to be overrepresented in shared selection outlier regions. This observation matches the commonly found differences and divergence in skin pigmentation between benthic and pelagic ecomorphs of Arctic charr (Alexander and Adams 2000), with benthic charr often having reduced melanin pigmentation due to often deeper depth distributions in lakes (Gardner et al. 1988; Adams et al. 1998; Fraser et al. 1998; Hooker et al. 2016). Positive selection and differentiation accompanied by a reduction of skin pigmentation would therefore resemble processes known from the evolution of cave organisms (Borowsky and Cohen 2013; Carlini and Fong 2017; Aardema et al. 2020). Even though cave organisms are an extreme case of low-light environmental adaptations, different lake environments might nevertheless favour specific skin pigmentations in Arctic charr that are likely under positive natural selection, as seen in other examples from fishes (e.g. Seehausen et al. 1999; Clarke and Schluter 2011; Wainwright 2019), which would explain the enrichment of melanosome functions we observed here. Assessments of melanosome function in relation to differences in genes under selection between the benthic and pelagic ecomorphs are necessary to establish the evolutionary significance of these findings.

Both in the benthic and pelagic ecomorphs, functions involved in the regulation of gene expression were overrepresented. This matches the common observation of the important role of gene regulation in many model systems of adaptive divergence (Mack and Nachman 2017; Singh et al. 2017; Franchini et al. 2019; Campana et al. 2020; Jacobs et al. 2020). How the selected regulatory genes affect the regulation of gene expression in nature remains to be tested using detailed RNA-sequencing analyses, ideally across larger parts of the species range of Arctic charr.

4.6 Conclusions

The repeated diversifications of Arctic charr are a prime example for replicated, parallel phenotypic evolution on a short time scale (Elmer and Meyer 2011; Klemetsen 2013; Elmer et al. 2014; Westram et al. 2014, 2016, 2018; Elmer 2016; Oke et al. 2017; Bolnick et al. 2018; Morales et al. 2019). We observed sharing of outlier genes across methods and benthic-pelagic ecomorph pairs as well as functional overrepresentation in shared outlier regions. Not only are the five shared outlier genes and large-scale outlier regions candidates for the study of repeated local adaptation, but they also highlight the ecological and functional challenges faced by diverging ecomorphs. Given the occurrence of the recovered outlier genes in other studies of local adaptation as well as their overlap with ecologically relevant salmonid QTLs, these genes may be relevant across other ecomorph divergences, not only in Arctic charr and salmonids. The high number of substitutions in the up- and downstream regions of outlier genes and the lack of changes in the coding regions together with the overrepresentation of gene expression-related functions argue for an important role of gene regulation in parallel divergences.

Chapter 5: Adaptive introgression after secondary contact facilitates ecomorph diversity in Arctic charr

5.1 Abstract

The salmonid fish Arctic charr (*Salvelinus alpinus*) is renowned for its abundant diversification in numerous northern freshwater lakes, where two or more ecomorphs occur in sympatry. In the case of Loch Dùghaill in Scotland, the benthic and pelagic ecomorphs are thought to have come into secondary contact after a time of allopatric separation of ancestral populations. Whole-genome resequencing data and the availability of a population sharing close ancestry with the pelagic ecomorph (Loch Uaine) enabled the application of four-taxon tests to elucidate introgression patterns across the genome between the two ecomorphs of Loch Dùghaill. Moreover, we used the genomic data to screen for targets of positive selection and, by determining the overlap of introgressed and selected regions, to identify adaptive introgression candidates and the relationship between selection and the extent of introgression. We found that positive selection in benthic charr was associated with stronger introgression, while positive selection in pelagic charr was associated with weaker introgression between the ecomorphs of Loch Dùghaill. This asymmetric introgression suggests gene flow of locally adapted alleles from one ecomorph to the other. Among the adaptive introgression candidates were genes potentially involved in memory formation, olfaction, food intake, growth, and age-at-maturity.

5.2 Introduction

When locally adapted populations of the same species come into contact, with or without preceding isolation in allopatry, their genomes are subject to the often conflicting processes of natural selection leading to adaptation and gene flow that potentially erodes adaptation (Lenormand 2002; Pedersen et al. 2017; Yamasaki et al. 2020). Nonetheless, the onset of gene flow after (secondary) contact does not necessarily lead to cessation of divergence, as reproductive barriers that arose through adaptation and are often linked to a reduction in local recombination rates can limit the erosive effects of gene flow (Nosil 2008; Lindtke and Buerkle 2015; Kulmuni et al. 2020). In some cases, introgression can even facilitate adaptation rather than hinder it (Pardo-Diaz et al. 2012; Jones et al. 2018; Ravinet et al. 2018). These cases of adaptive introgression are thought to be a catalyst to repeated adaptation to the same environment (Bay et al. 2019; Giska et al. 2019; Marques et al. 2019) and have been evidenced in diverse systems, from human-Neanderthal to

cichlids to butterflies (Pardo-Diaz et al. 2012; Enciso-Romero et al. 2017; Racimo et al. 2017; Malinsky et al. 2018).

So far, the interplay between selection and introgression at the whole-genome level has only been studied in a small number of systems (Richards et al. 2018; Walsh et al. 2018; Martin et al. 2019; Yamasaki et al. 2020), partly owing to the elusive nature of genomic signatures of the speciation process, which leads to dramatically reduced power to detect natural selection and gene flow as time passes and subsequent demographic processes obscure signals (Ravinet et al. 2017; Schneider et al. 2020). By studying recent or ongoing adaptive divergences with gene flow, those challenges can be circumvented. A study system for recent adaptive divergences that has recently received growing attention are the diversifications of Arctic charr (*Salvelinus alpinus*) in postglacial lakes of the northern hemisphere (Elmer 2016), also owing to the high phenotypic parallelism of repeated divergences into ecomorphs (Jacobs et al. 2020). One case of recent diversification in Arctic charr that is particularly suitable for studying the interplay between selection and introgression is Loch Dùghaill, containing one benthic (profundal, benthivorous) and one pelagic (limnetic, planktivorous) ecomorph, the latter of which is more closely related to a single pelagic ecomorph in the neighbouring Loch Uaine (hereafter called 'allopatric pelagic') than to the benthic ecomorph in Loch Dùghaill (Hooker et al. 2016; Jacobs et al. 2020). These two ecomorphs have distinct morphologies and colourations, and occupy different depths of the lakes displaying different feeding behaviours (Pigeon et al. 1997; Alexander and Adams 2000; Elmer 2016; Knudsen et al. 2016; Jacobs et al. 2020) in ways that are stereotypical of morph divergences in freshwater lakes (Cooper et al. 2010; Matthews et al. 2010; Lundsgaard-Hansen et al. 2013). Evidence from RAD-sequencing data and demographic modelling suggests that the two ecomorphs in Loch Dùghaill came into secondary contact after having diverged an estimated 50 Kya, even though continuing divergence is possible (Jacobs et al. 2020). There is also evidence for strong ongoing gene flow between the ecomorphs of Loch Dùghaill and evidence for response to selection shaping divergence at targetted genomic regions (Jacobs et al. 2020). However, it is unknown to what extent gene flow has an influence on the existing genomic adaptation in the two ecomorphs. Moreover, there is also the possibility of adaptive introgression, especially from the ecomorph that colonised the lake first and therefore is expected to have had more opportunity to evolve local adaptation to the lake conditions, even though it is unclear which ecomorph that was (Jacobs et al. 2020). Lastly, we do not know which specific genomic regions are under selection in this remarkable case of adaptive divergence.

Here, we use whole-genome resequencing to address the gaps in knowledge about the diversification of Arctic charr as a case study for divergence after (secondary) contact with potentially ongoing gene flow typical of postglacial lakes, here from two ecomorphs in Loch Dùghaill and the putative source population of the pelagic ecomorph, the allopatric pelagic Loch Uaine charr (Jacobs et al. 2020). Specifically, we explore the role of adaptive introgression after secondary contact in the diversification and maintenance of new biodiversity in the lake, in the form of ecologically distinct and reproductively isolated benthic and pelagic ecomorphs. The flux of water levels and population connectivity in temperate regions during glacial cycles means these are likely pervasive yet to date unexamined genomic mechanisms of diversification in postglacial fishes (Schluter 1996; Brunner et al. 2001; Esin et al. 2017). We 1) analyse genomic targets of selection using a powerful recently developed approach that makes use of three different properties of the genome to infer signals of positive selection, calculating a statistic called μ that is based on the joint signal of the site frequency spectrum, linkage disequilibrium, and a reduction in local diversity (Alachiotis and Pavlidis 2018), and 2) investigate the overlap of signals of selection with signals of introgression inferred using the patterns of genetic variants in (sub)population quartets (Patterson et al. 2012; Ward and Van Oosterhout 2016). In step 3), we analyse the gene content and functional enrichment in regions under selection, introgression, and both selection and introgression together. This allowed us to assess the presence of adaptive introgression, the associated genes, and their functional significance to tell us more about how adaptively introgressed genes might influence local adaptation and the diversification process.

We found melanosome-related functions to be enriched in selection outliers of the benthic ecomorph, while more diverse functions were overrepresented in the pelagic ecomorph. Among the adaptive introgression candidates with selection in and introgression between ecomorphs, we identified genes potentially involved in memory formation, olfaction, food intake, growth, and age-at-maturity. Our whole-genome data indicate that regions under selection in the benthic ecomorph show stronger signs of introgression than those in the pelagic ecomorph, which could indicate that gene flow of locally adaptive variants from the first coloniser of the lake occurred.

5.3 Materials & methods

5.3.1 Library preparation & whole-genome sequencing

We used NEBNext Ultra II FS DNA Library Prep Kit for Illumina (New England BioLab) to generate libraries for 26 individuals from Loch Dùghaill (13 benthic and 13 pelagic fish) and four individuals from the allopatric pelagic charr in Loch Uaine (a loch with a charr population that has a shared demographic history with Loch Dùghaill, Jacobs et al. 2020) for a total of 30 individuals (sampled populations: Figure 5.1). The sampled individuals included males and females in equal proportions. See Chapter 4 for details on DNA extraction and library preparation. The whole-genome libraries were sequenced at a coverage of ~5X per individual (based on 3 Gb genome size) on the Illumina HiSeqX Ten platform at BGI, Hong Kong, on five lanes using 2x150 bp reads.

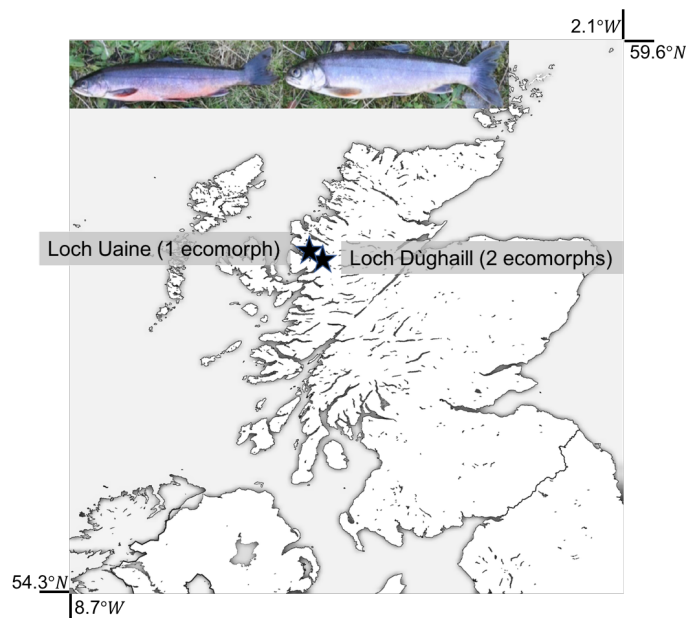


Figure 5.1 – Sampling locations of Arctic charr. The stars indicate the geographic position of the lakes. Loch Uaine contains an allopatric, closely related pelagic ecomorph. The number of included ecomorphs are shown in parentheses. In the top left, individuals of the pelagic ecomorph (left) and benthic ecomorph (right) from Loch Dùghaill are shown (fish images from Hooker et al. 2016). Also note the colour differences between the ecomorphs. Map: Creative Commons License (CC BY 2.5), www.demis.nl

5.3.2 Genotype calling, SNP calling, & SNP filtering

The complete data set was quality-checked with FastQC (Andrews 2010), SAMtools (Li et al. 2009), and Picard (<http://broadinstitute.github.io/picard>). Reads were filtered and library adaptors were removed with Trimmomatic (Bolger et al. 2014) using an ILLUMINACLIP seed mismatch threshold of 2, a palindrome clip threshold of 30, a simple clip threshold of 10, and a minimum adaptor length of 7. Leading and trailing bases with a PHRED score below 20 were removed. A 4-bp sliding window approach was used to remove reads that contained windows with a PHRED score below 20. A minimum read length of 30 bp was specified in Trimmomatic. Remaining adaptor remnants identified using FastQC were removed using cutadapt with an error tolerance of 0 (Martin 2011). After filtering, the cleaned reads were then again checked using FastQC to ensure that median read quality PHRED scores were above 30 and to confirm the absence of adaptors and sequencing artefacts. The filtered and cleaned reads were then mapped against the Arctic charr reference genome version GCF_002910315.2_ASM291031v2 (Christensen et al. 2018) using the mem algorithm of BWA (Li and Durbin 2010). After sorting of BAM files with SAMtools, deduplication using GATK4.0.8.1 (McKenna et al. 2010), sequence library creation using Picard, and insertion-deletion (INDEL) target creation using all individuals as well as INDEL realignment using GATK3.8.1, we used GATK4.0.8.1 for genotype and SNP calling.

We followed the developers' recommendations for genotype calling, SNP calling, and SNP filtering (<https://gatk.broadinstitute.org/hc/enus/articles/360035531112?id=23216>). First, we used HaplotypeCaller on the indexed BAM files separately for each individual. Next, we combined genotype vcf (gvcf) files and genotyped the combined individuals in GATK4.0.8.1. For genotyping, we provided an expected heterozygosity value of 0.0007, which was calculated using Picard BamIndexStats (<http://broadinstitute.github.io/picard>). For filtering, we followed the GATK4 developers' recommendations (<https://gatk.broadinstitute.org/hc/enus/articles/360035531112?id=23216>): variant quality/depth of alternate allele (QualByDepth) > 2, PHRED-scaled P values of Fisher's Exact Tests for strand bias (FisherStrand) < 60, Root Mean Square of mapping quality across samples (RMSMappingQuality) > 40, U-based z-approximation from Mann-Whitney Rank Sum Test for mapping qualities (MappingQualityRankSumTest) > -12.5, U-based z-approximation from Mann-Whitney Rank Sum Test for distance from end of reads for reads with an alternate allele (ReadPosRankSumTest) > -8, strand bias odds ratio (StrandOddsRatio) < 3. Due to the relatively low coverage of the data, we additionally applied two further filters: minor allele frequency (MAF) > 0.05 and, to remove variants deviating from the Hardy-Weinberg Equilibrium, excess heterozygosity (ExcessHet) < 30.

Insertions and deletions (INDELs) were removed for downstream analyses. Filtered SNP variants were then phased with BEAGLE (Browning and Browning 2007) to obtain haplotype information.

5.3.3 Population structure & four-taxon tests for introgression

To elucidate the population structure of Loch Dùghaill and the allopatric pelagic ecomorph, we used the PCA-based approach SNPRelate on the final SNP data set (Zheng et al. 2012, 2017). We used the first two principal components to describe population structure, as the amount of genetic variation explained dropped after the first principal component. Using the first four PC scores, we then plotted the PCAngsd results using a custom R script in R v3.5.1 (R Core Team 2018).

Patterns of admixture and introgression among individuals within and across populations were inferred in Chapter 4 using NGSadmix (Skotte et al. 2013). To estimate introgression between the sympatric benthic and pelagic ecomorphs of Loch Dùghaill, we performed four-taxon tests in HybridCheck v1.0.1 on all biallelic SNPs (Ward and Van Oosterhout 2016). We used a block length of 200 Kbp for the sliding window analysis across the genome. *D* statistics were calculated in such a way that positive values indicated introgression between the benthic and pelagic ecomorphs of Loch Dùghaill (ABBA pattern), and negative values indicated introgression between Dùghaill benthic and allopatric pelagic (BABA pattern) (Figure 5.2). We defined windows with strong evidence for introgression between the ecomorphs of Loch Dùghaill as those exceeding the 99.5% quantile of the *D* statistic.

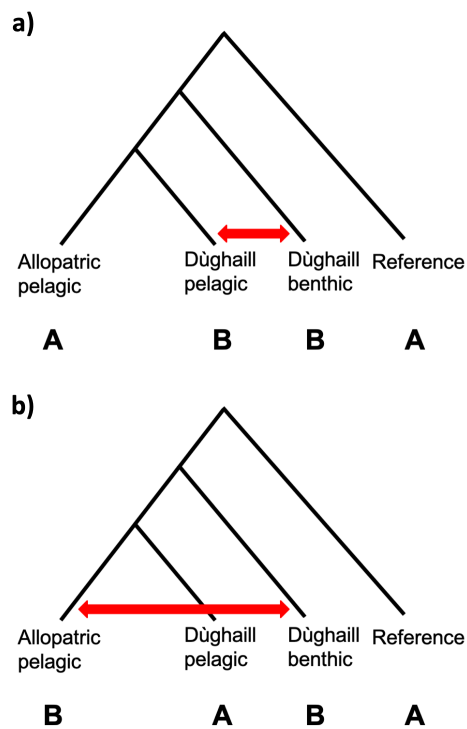


Figure 5.2 – Schematics of two introgression patterns tested in a four-taxon test. a) ABBA pattern more common: shared derived variants in benthic and pelagic charr suggesting introgression between ecomorphs of Loch Dùghaill. b) BABA pattern more common: shared derived variants in allopatric pelagic and benthic Dùghaill charr suggesting ancient introgression, retained ancestral polymorphism, or incomplete lineage sorting.

5.3.4 Selection scans

Scans for signals of positive selection were performed using RAI SD v2.9 (Alachiotis and Pavlidis 2018), which calculates the μ statistic based on summarised SNP vectors. The μ statistic collectively includes information from 1) the reduction of local variation (π), 2) shifts in the site frequency spectrum (SFS; comparison of high- and low-frequency derived variants), and 3) localised linkage disequilibrium (LD) patterns. The μ statistic was calculated based on sliding windows consisting of 50 SNPs each. We defined windows under positive selection as those exceeding the 99.5% quantile of the μ statistic.

5.3.5 Inference of adaptive introgression

We defined regions under putative adaptive introgression as those having strong evidence for positive selection (μ statistic above 99.5% quantile) in either the benthic or pelagic ecomorph (or both) and strong evidence for either introgression between the benthic and pelagic ecomorphs in Loch Dùghaill (D statistic exceeding 99.5% quantile) or between the benthic ecomorph of Loch Dùghaill and allopatric pelagic charr (D statistic below 0.5%

quantile). The overlap of windows under positive selection and introgression was determined using custom R scripts using the R package IRanges v2.16.0 (Lawrence et al. 2013) to assess the physical overlap of outlier window positions.

To infer whether positive selection in Loch Dùghaill charr is associated with a shift in the D statistic, we 1) tested for deviations of the D statistic in windows exceeding the 95% or 99.5% quantile of the μ statistic in the benthic or pelagic ecomorph of Loch Dùghaill and 2) used correlation analysis of the μ and D statistics, separately for the benthic and pelagic ecomorphs of Loch Dùghaill. Both association analyses were performed in R v3.5.1 (R Core Team 2018).

5.3.6 Functional enrichment

For gene ontology (GO) term enrichment analyses in both the RAI_{SD} and four-taxon test outlier sets (RAI_{SD}: above 99.5% quantile; four-taxon test D statistic: above 99.5% quantile or below 0.5% quantile), we first extracted zebrafish gene symbols of all outlier genes from the Arctic charr reference genome annotation file (Christensen et al. 2018). We converted part of the gene symbols for which only the locus number and gene name but not the gene symbol were present in the annotation file based on the MGI vertebrate homology database (<http://www.informatics.jax.org/homology.shtml>) and UniProt/SwissProt gene information. A GO term overrepresentation analysis on the gene symbols was performed in WebGestalt v2019 (<http://www.webgestalt.org/>; last access: 21.09.2020) (Zhang et al. 2005; Wang et al. 2013, 2017; Liao et al. 2019) with a zebrafish GO term reference database. For this, we used biological process GO terms and the following analysis parameters: 1) a minimum number of genes per category of 3, 2) a maximum number of genes per category of 2000, 3) a Benjamini-Hochberg FDR correction (Benjamini and Hochberg 1995), 4) retaining the top 10 overrepresented GO terms, 5) a number of categories expected from set cover of 10, and 6) a number of visualised categories of 40. GO terms with FDR q values below 0.05 were deemed significantly overrepresented, while those with P values below 0.05 but FDR q values above 0.05 were considered to show a trend of overrepresentation. We used these two categories of overrepresentation to avoid being too conservative in our inference, since many of the GO terms are only represented by a small number of genes.

5.4 Results

5.4.1 Mapping success and number of SNPs

The average coverage of Loch Dùghaill and the allopatric pelagic ecomorph was 4.95X and 4.32X (calculated using covered and uncovered regions based on a genome size of 3 Gb), respectively. Regions of the genome with at least 1X coverage had an average coverage of 10.26X (Dùghaill) and 8.77X (allopatric pelagic). The total number of SNPs after filtering was 13,853,994.

5.4.2 Population structure

The PCA of population structure revealed a complete separation of the allopatric pelagic ecomorph and each of the ecomorphs in Loch Dùghaill along PC1 and a more subtle separation of ecomorphs in Loch Dùghaill and the allopatric pelagic ecomorph along PC2 (Figure 5.3). PC1 explained 14.66% of genetic variation, while PC2 explained 3.81%. One individual each of the benthic and pelagic ecomorphs showed signs consistent with pronounced introgression or hybridisation, as they approached the ecomorph clusters along PC1 (Figure 5.3 and admixture plot in Chapter 4: Figure 4.3).

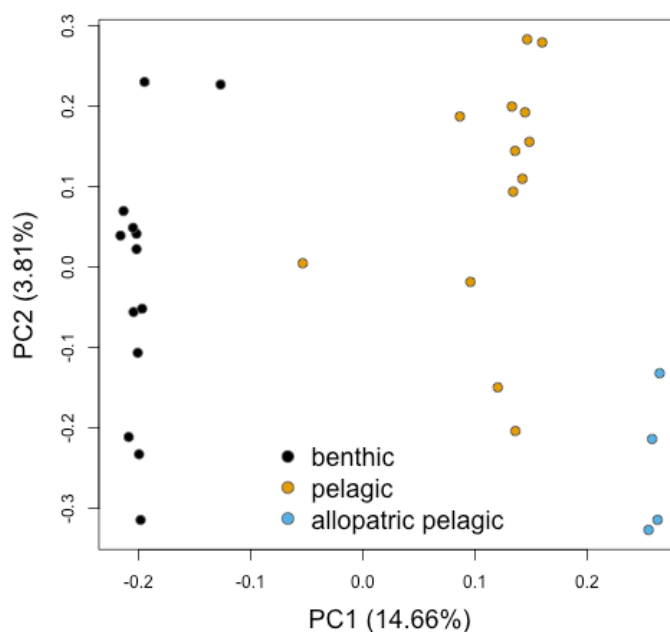


Figure 5.3 – SNP-Relate population structure plot of PC1 against PC2. The percentages in parentheses represent the variance explained by the two PCs. benthic = benthic Loch Dùghaill ecomorph; pelagic = pelagic Loch Dùghaill ecomorph; allopatric pelagic = allopatric pelagic Loch Uaine ecomorph

5.4.3 Gene functions under selection

Regions under selection were spread relatively evenly across the genome (Figure 5.4a,b). The outlier windows with evidence for positive selection (μ statistic > 99.5% quantile) differed between the benthic and pelagic ecomorphs of Loch Dùghaill (Figure 5.4a,b). This was also evident in the functional enrichment analysis in selection outlier windows of benthic Loch Dùghaill charr, which mainly revealed GO terms related to melanosome function and transport as overrepresented (Fisher's Exact Tests: $P < 0.05$; Figure 5.5a). In RAiSD outlier windows of pelagic Loch Dùghaill charr, a more diverse set of GO terms was overrepresented, including mitochondrial, nucleic acid metabolism, immunity, and lipid transport and metabolism functions (Fisher's Exact Tests: $P < 0.05$; Figure 5.5b).

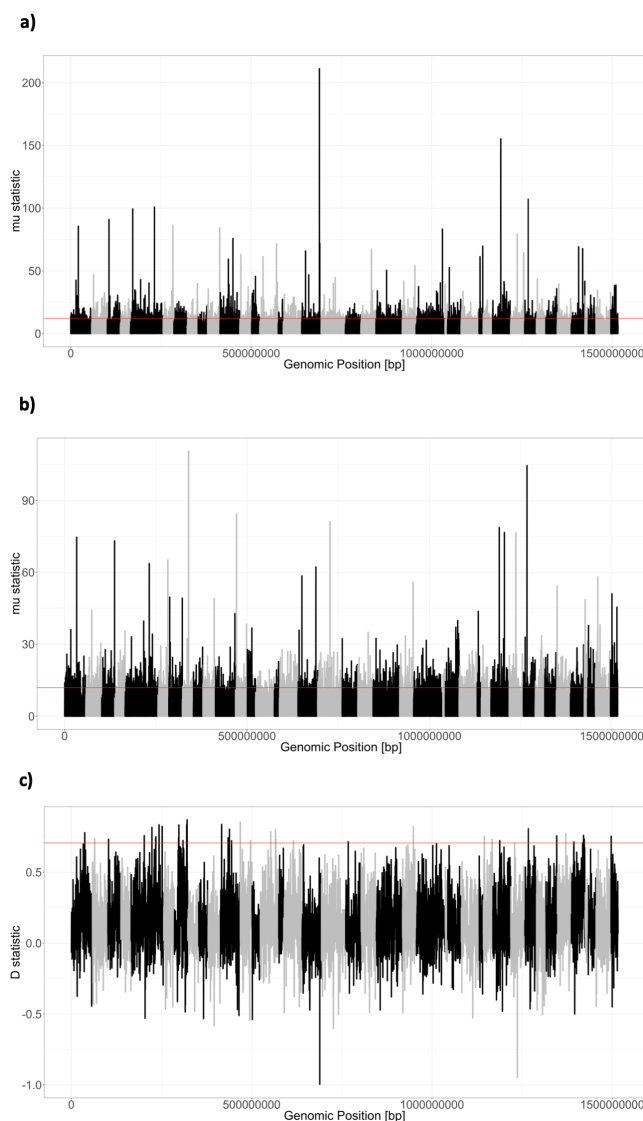
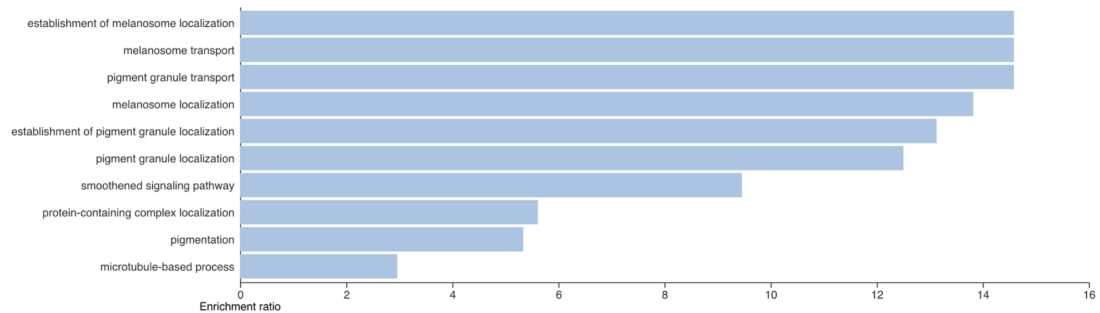


Figure 5.4 – Manhattan plots of the selection statistic μ and the introgression statistic D across the genome in Loch Dùghaill charr calculated using RAiSD and HybridCheck, respectively. a) μ statistic in benthic Loch Dùghaill ecomorph; b) μ statistic in pelagic

Loch Dùghaill ecomorph; c) D statistic, with positive values indicating introgression between the benthic and pelagic ecomorphs of Loch Dùghaill (ABBA pattern), and negative values indicating introgression between Dùghaill benthic and the allopatric pelagic ecomorph (BABA pattern). The red horizontal bars indicate the 99.5% quantiles of the μ and D statistics. The black and white shadings indicate the assembled chromosomes of the Arctic charr genome.

a)



b)

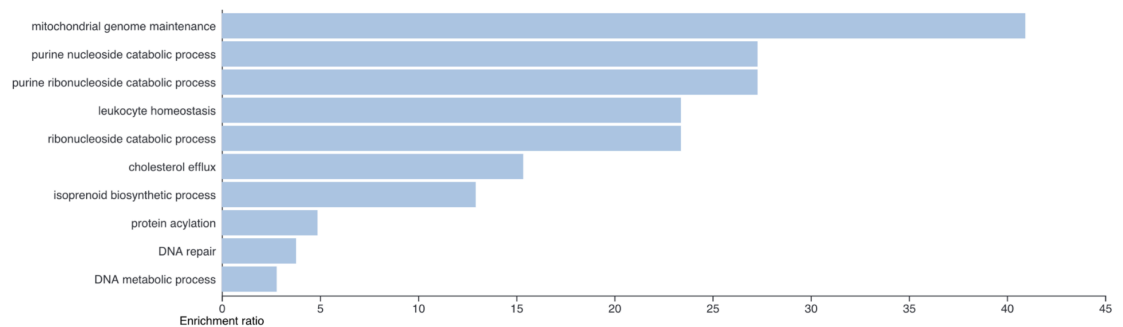


Figure 5.5 – Top 10 overrepresented gene categories in genes under positive selection. a) Overrepresented gene categories (GO terms) in genes under selection in the benthic ecomorph. b) Overrepresented gene categories (GO terms) in genes under selection in the pelagic ecomorph. The x-axis shows the ratio of observed to expected frequencies of GO terms.

5.4.4 Introgression between the benthic & pelagic ecomorphs

The distribution of the D statistic in our four-taxon tests of introgression showed a pronounced shift to positive values (global average of $D = 0.144$), indicating pervasive introgression between the benthic and pelagic ecomorphs of Loch Dùghaill (Figure 5.4c). Only a minority of windows were shifted to negative values.

5.4.5 Regions of putative adaptive introgression between ecomorphs

In regions under positive selection ($\mu > 95\%$ quantile) and regions under strong positive selection ($\mu > 99.5\%$ quantile) in the benthic ecomorph, the window-based four-taxon test revealed a shift towards stronger introgression (higher D) between the benthic and pelagic ecomorphs of Loch Dùghaill ($D = 0.1654$ and 0.1945 , respectively, vs $D_{\text{global}} = 0.1445$; Table 5.1). For regions under positive selection ($\mu > 95\%$ quantile) and regions under strong positive selection ($\mu > 99.5\%$ quantile) in the pelagic ecomorph of Loch Dùghaill, the opposite was true, i.e. we observed a shift towards weaker introgression (lower D) between the benthic and pelagic ecomorphs of Loch Dùghaill, while the average amount of introgression was still positive ($D = 0.1327$ and 0.1347 , respectively, vs $D_{\text{global}} = 0.1445$; Table 5.1).

Table 5.1 – Shifts in introgression in regions under positive selection. Differences in average D statistic values between global (all) regions and regions under positive selection in benthic and/or pelagic charr according to RAI_{SD} results are shown along with associated t-test results and conclusions about introgression within Loch Dùghaill.

Selection set	Difference (D)	Global (D)	Selected (D)	t statistic	P	Conclusion
benthic (95% quantile)	-0.021	0.145	0.165	-9.375	8.86E-21	Stronger introgression between Dùghaill ecomorphs
benthic (99.5% quantile)	-0.050	0.145	0.195	-19.446	3.94E-83	Stronger introgression between Dùghaill ecomorphs
pelagic (95% quantile)	0.012	0.145	0.133	5.344	9.33E-08	Weaker introgression between Dùghaill ecomorphs
pelagic (99.5% quantile)	0.010	0.145	0.135	3.900	9.68E-05	Weaker introgression between Dùghaill ecomorphs

In benthic charr we observed a positive correlation between the estimate of positive selection (μ) and the estimate of introgression (D) (Pearson's correlation test: $r = 0.0318$, $t = 5.5330$, $P = 3.1737\text{E-}08$), indicating elevated introgression between the ecomorphs occurring at genomic regions under positive selection in the benthic ecomorph. In contrast, in the pelagic ecomorph, the correlation between positive selection and the D statistic was negative (Pearson's correlation test: $r = -0.0417$, $t = -7.2754$, $P = 3.5368\text{E-}13$), indicating

weaker introgression between the benthic and pelagic ecomorphs in genomic regions that are under positive selection in the pelagic ecomorph.

5.4.6 Adaptive introgression candidate genes

The regions under putative adaptive introgression with positive selection in benthic charr contained 29 genes (24 protein-coding genes, 2 long non-coding RNAs, and 3 pseudogenes), such as the neural genes *tenascin*, *protein turtle homolog A*, and *regulator of G-protein signalling 14* as well as the developmental gene *protein diaphanous homolog 2* (Table S5.1). In pelagic charr, the regions under putative introgression and positive selection contained 18 genes (16 protein-coding genes and 2 pseudogenes, such as *low-density lipoprotein receptor-related protein 11*, *SRSF protein kinase 3*, and *pseudouridylate synthase 7 homolog* (Table S5.2). Lastly, the genes with a signal of introgression and under positive selection in both benthic and pelagic charr, i.e. those which are expected to be adaptive in both ecomorphs, included 12 genes, seven of which were characterised coding genes, two were uncharacterised coding genes, and three were pseudogenes (Table 2; Table S5.3). Among the characterised coding genes were *trace amine-associated receptor 1*, *protein phosphatase 1L*, *ankyrin repeat domain-containing protein 34C*, and *serine/threonine-protein kinase LATSL*.

Regions under both strong positive selection in the benthic ecomorph ($\mu > 99.5\%$ quantile) and subject to introgression between the benthic and pelagic Loch Dùghaill ecomorphs ($D > 99.5\%$ quantile) were overrepresented for genes with functions involved in neural processes, immunity, development, and response to stress (Fisher's Exact Tests: $P < 0.05$; Figure 5.6a). In contrast, regions under both strong positive selection in the pelagic ecomorph ($\mu > 99.5\%$ quantile) and subject to elevated levels of introgression between ecomorphs ($D > 99.5\%$ quantile) were overrepresented for genes involved in catecholamine metabolic processes, organ growth, microtubule severing, hippo signalling, and pseudouridine synthesis (Fisher's Exact Tests: $P < 0.05$; Figure 5.6b).

Table 5.2 – List of adaptive introgression candidate genes. These genes were found to be under positive selection in both benthic and pelagic charr and showed signals of introgression between these two ecomorphs of Loch Dùghaill. Class = class of the gene; uncharacterised = unknown product or protein name of coding gene; μ = measure of positive selection in RAI SD software; D = measure of introgression between benthic and pelagic ecomorphs in Loch Dùghaill.

Chromosome	Start	End	Class	Gene symbol	mu	D
NC_036842	36628924	36685678	pseudogene	–	12.38	0.758
NC_036842	36635290	36695860	pseudogene	–	14.89	0.758
NC_036842	68870162	68893196	gene	ankrd34c	12.56	0.749
NC_036848	30878936	30907308	uncharacterised	–	14.41	0.722
NC_036848	30989907	31006263	uncharacterised	–	12.98	0.722
NC_036859	29996030	30054180	gene	ppm1l	12.23	0.822
NC_036868	9784845	9891929	gene	lats1	19.44	0.807
NC_036868	9801202	10129317	gene	nup43	72.24	0.807
NC_036868	9807474	10130095	gene	pcmt	85.75	0.807
NC_036868	9913810	10142213	gene	moxd1	48.56	0.807
NC_036868	10027253	10142256	pseudogene	–	19.72	0.807
NC_036868	10030187	10142261	gene	taar1	13.98	0.807

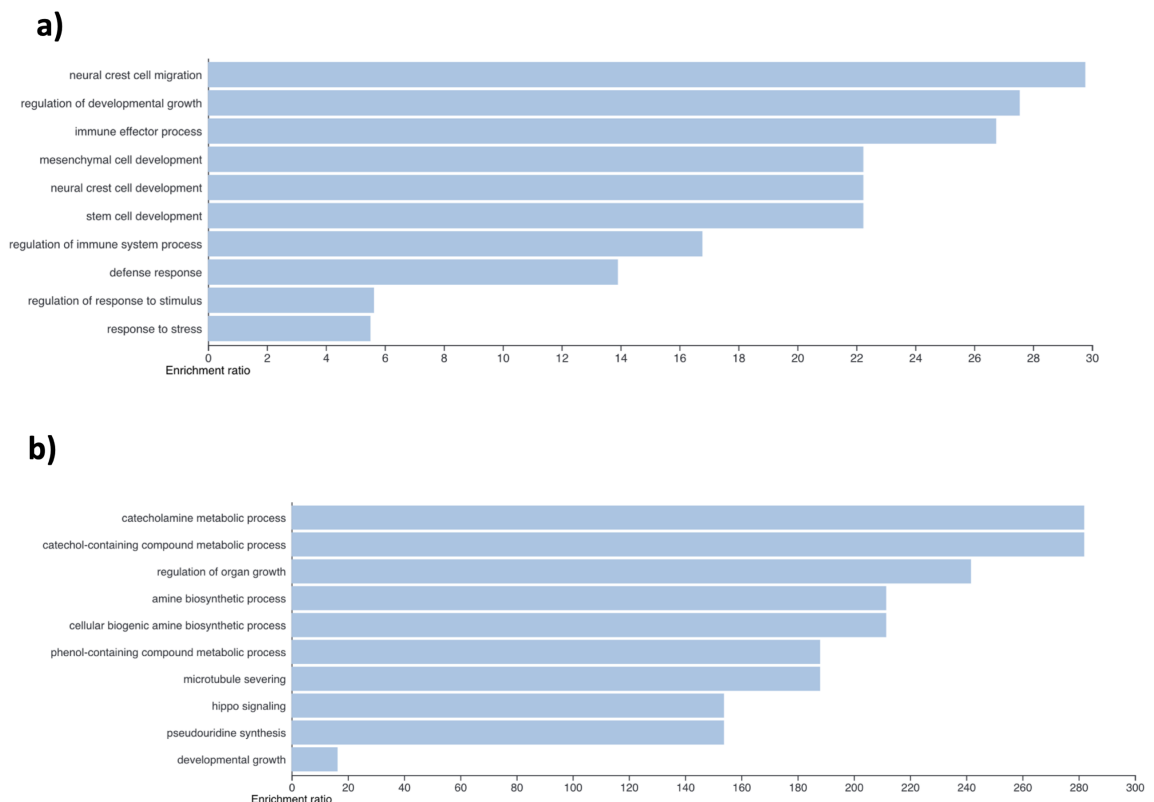


Figure 5.6 – Top 10 overrepresented gene categories in genes under positive selection and subject to introgression between ecomorphs. a) Overrepresented gene categories (GO terms) in genes under selection in the benthic ecomorph and introgression between benthic and pelagic ecomorphs. b) Overrepresented gene categories (GO terms) in genes under selection in the pelagic ecomorph and introgression between benthic and pelagic ecomorphs. The x-axis shows the ratio of observed to expected frequencies of GO terms.

5.5 Discussion

Our analyses of positive selection and introgression in a case study of rapidly diversifying benthic and pelagic Arctic charr ecomorphs and an allopatric, closely related population suggest: 1) the extent of introgression between the ecomorphs of Loch Dùghaill is different in regions under positive selection in benthic charr compared to pelagic charr; 2) there is a positive relationship between the extent of positive selection in benthic charr and the magnitude of introgression between ecomorphs; 3) protein-coding genes under selection in both ecomorphs introgressed from one to the other ecomorph. These findings indicate that gene flow between ecomorphs is not random but rather can preferentially include regions that are under positive selection in both ecomorphs within a lake. Moreover, we found that the functional targets of positive selection differ between ecomorphs, reflecting their differing role in phenotype. Our study contributes to the novel insights about the mechanisms behind and origins of functional adaptive variation.

5.5.1 Gene functions under selection & adaptive role

We observed that melanosome-related functions were highly overrepresented in the benthic ecomorph but not in the pelagic ecomorph, which is an indication of differential selection pressures in these melanic pigment cells. This result agrees with the deeper-water distribution of the benthic (or profundal) ecomorph, which has also been found to have reduced skin pigmentation relative to the pelagic ecomorph (Alexander and Adams 2000; Hooker et al. 2016). Such reduced skin pigmentation has also been found in other benthic ecomorphs of Arctic charr (Gardner et al. 1988; Adams et al. 1998; Fraser et al. 1998).

On the other hand, gene functions overrepresented in the pelagic ecomorph were of a more diverse nature. Lipid transport and metabolism have been found to be under selection and/or differential gene regulation in studies of Arctic charr (Schneider et al. 2019) as well as other salmonids (Baillie et al. 2016; Gillard et al. 2018), which has been attributed to the importance of lipid metabolism for life history as well as known buoyancy differences regulating depth optima between ecomorphs (Jobling et al. 1998). Immunity genes have been established as targets of positive or diversifying selection in salmonids (Pritchard et al. 2018; Schneider et al. 2019) and various model systems of ecological divergence (Anaya-Rojas et al. 2016; Stutz and Bolnick 2017). Mitochondrial function has also been found to be under selection in some systems of salmonids (Garvin et al. 2011). Together, these overrepresented functions point towards potentially important adaptations needed for ecological adaptation and adaptive diversification in salmonids and beyond.

5.5.2 Relationship between positive selection and introgression

The elevated introgression we observed for regions under positive selection in benthic charr and a lower amount of introgression in those under selection in pelagic charr suggests the likely directionality of gene flow in the evolutionary history of ecomorph divergence. A demographic history of secondary contact that was inferred for Loch Dùghaill charr (Jacobs et al. 2020) and the close relationship to the allopatric pelagic ecomorph, which inhabits a small lake on a hill northwest of the valley where Loch Dùghaill is located, to pelagic Loch Dùghaill charr indicates the colonisation of Loch Dùghaill was likely by pelagic charr from, or simultaneous to, the allopatric pelagic ecomorph of Loch Uaine. The lines of evidence suggest that the ancestral version of the benthic ecomorph, on the other hand, colonised Loch Dùghaill before the pelagic ecomorph, after which it came into contact with the newly arrived pelagic ecomorph. This hypothetical scenario would explain the higher amount of introgression in regions of positive selection in the benthic ecomorph, as a consequence of it having a more advanced level of local adaptation compared to the subsequent pelagic coloniser (Bisschop et al. 2019; Germino et al. 2019). Locally adapted variants could then have introgressed from the benthic to the pelagic ecomorph, similar to the interspecific adaptive introgression observed in cichlids (Malinsky et al. 2018), butterflies (Pardo-Diaz et al. 2012), humans and Neanderthals (Enciso-Romero et al. 2017; Racimo et al. 2017), and even in some plant groups, such as oaks (*Quercus* sp.) (Leroy et al. 2019). However, the exact colonisation and contact scenarios, in particular the directionality of gene flow, remain to be tested explicitly, for example using demographic modelling and comparisons to more neighbouring populations.

Relative to genome-wide levels, we found lower levels of introgression between ecomorphs in genomic regions that were under positive selection in pelagic charr, not only compared to regions under selection in benthic charr but also compared to the global average. This suggests that these regions are more resistant to gene flow than other genomic regions, which could be explained by their important adaptive role for pelagic charr (Butlin et al. 2014; Ravinet et al. 2017). Such cases of genomic regions with higher resistance to gene flow than the remainder of the genome, also called genomic islands of divergence, have been observed in a plethora of biological systems (Nadeau et al. 2012; Renaut et al. 2013; Papadopoulos et al. 2019; Rodríguez-Ramilo et al. 2019; Rautsaw et al. 2020). The genes they contain have putatively high functional relevance for the process of ecological adaptation.

5.5.3 Potential adaptive introgression targets in ecological speciation of postglacial fishes

Genomic regions and genes that are subject to introgression and are under positive selection in both ecomorphs of Loch Dùghaill are candidates for adaptive introgression, as they play a putative adaptive role in both ecomorphs, irrespective of the directionality of gene flow. This kind of adaptive gene flow has been shown to contribute to ecological speciation and diversification, such as the adaptive introgression observed in microendemic radiations of pupfishes in the Bahamas (Richards and Martin 2017).

Adaptive introgression has found increasing attention in recent years and is now thought to be an important catalyst facilitating ecological speciation, for example in trout (Jacobs et al. 2018), sticklebacks (Marques et al. 2019), *Heliconius* butterflies (Pardo-Diaz et al. 2012), and cichlids (Malinsky et al. 2018). Seven characterised protein-coding genes were among the putatively adaptively introgressed genes, including a gene involved in olfaction, food intake, and memory formation (*taar1*) (Rutigliano et al. 2018). *Taar1* belongs to a family of genes that has been found to be important for homing behaviour in Atlantic salmon (*Salmo salar*) (Tessarolo et al. 2014). Another adaptive introgression candidate, *lats1*, has been associated with age-at-maturity in Atlantic salmon (Kurko et al. 2020). Lastly, the putatively adaptively introgressed gene *ankrd34c* has been implicated in body growth in bighead carp (Liu et al. 2016).

It remains to be tested in which direction these putative targets of adaptive introgression flow (from benthic to pelagic or vice versa) and whether they occur in regions of low recombination, as for example has been observed for locally adapted genes in an *Arabidopsis lyrata* model system (Hämälä and Savolainen 2019). Nevertheless, the robustly supported candidate genes and regions that we identified in these ecomorphs of Arctic charr constitute novel evidence for the role of adaptive introgression in facilitating the rapid adaptive divergence of Arctic charr and add to the evidence of the role of adaptive introgression in rapid diversifications and ecological speciation.

Chapter 6: General discussion

6.1 Summary of thesis chapters

In the above thesis chapters I explored the sharing and similarities of natural selection at different levels, from transcriptomes across salmonids to whole genomes across populations of Arctic charr. I examined the influence of demographic history, migration, and genomic properties on screens for selective sweeps using forward-in-time simulations. In the final data chapter, I analysed the interplay of positive selection and introgression to infer putative adaptive introgression between sympatric ecomorphs in an exemplary population of Arctic charr.

One of the main results that runs like a common thread through Chapters 2 and 4 is that sharing of the targets of natural selection is limited to only a relatively small number of genes, while larger genomic regions are more commonly under shared selection. However, while these shared outlier genes are few, they are nevertheless putatively important for rapid diversification in salmonids and, judging from other studies that detected the same outliers, also for other taxonomic groups across the Metazoan kingdom. In Chapter 4, the putative ecological importance is underlined by the overlap of shared selection outlier genes with ecologically relevant QTLs. In terms of functions, the top outliers for both the transcriptomic data in Chapter 2 and the whole-genome data in Chapters 4 and 5 were commonly genes involved in neural processes, memory formation, and the regulation of feeding behaviour, which matches the observation of an overlap with QTLs involved in body size and growth. Indeed, many of the selected genes shared across Arctic charr populations are themselves linked to QTLs associated with feeding and growth in other taxonomic groups, particularly in fish. This result is in concordance with one of the main commonalities of phenotype in the same ecomorphs across populations being their diet, either benthic (benthivorous) or pelagic (planktivorous) (Pigeon et al. 1997; Elmer 2016; Knudsen et al. 2016; Jacobs et al. 2020), and, more generally, the importance of neural functions for rapid diversifications (Brawand et al. 2014; York et al. 2019; Liang et al. 2020).

Broader gene functions in shared genomic regions under selection in both Chapter 2 and Chapter 4 were overrepresented for the regulation of gene expression, which is in line with the pronounced parallelism of gene regulation found in whitefish (Derome et al. 2006) and charr (Jacobs et al. 2020). Moreover, both in Chapter 4 and Chapter 5, melanosome function and transport were among the most highly overrepresented gene functions in genomic regions under positive selection, which shows that the enrichment of

selected gene functions is similar at the across-population and the within-population level. In the interspecies comparison using transcriptomes in Chapter 2 there was, however, a higher degree of overrepresentation of muscle function and immunity relative to the across-population comparison in Arctic charr, which supports the notion that these functions could be necessary adaptations that allow rapid adaptation and diversification in charr and whitefish, which is in line with evidence from 66 teleost genomes showing a role of immunity genes in increased speciation rates (Malmstrøm et al. 2016) and the important contribution of muscle function in ecological diversifications of pupfishes (*Cyprinodon*) on Caribbean islands (Richards and Martin 2017).

Chapter 5 showed that many of the genes under positive selection were at the same time subject to introgression between sympatric ecomorphs. Moreover, about one half of the genes under both selection and introgression in benthic charr was also under selection and introgression in pelagic charr. These genes being under positive selection in both ecomorphs and showing evidence for introgression between them can be considered as candidates for adaptive introgression, putatively conferring increased local adaptation by the exchange of selected gene variants. The fact that adaptive introgression candidates included immunity and stress response genes supports a role of these functions in lake-specific rather than ecomorph-specific local adaptation, potentially due to lake-specific parasite and pathogen pressures (Douda et al. 2017) as well as lake-specific environmental variables, such as chemical composition and temperature (Kavanagh et al. 2010; Alcántara-Rodríguez et al. 2012; Franch-Gras et al. 2018). It is also possible that introgression in such regions leads to the exchange of variants for local adaptation to ecomorph-specific habitats from pools of standing genetic variation (Feulner et al. 2013; Marques et al. 2018; Urban et al. 2020).

The simulation study in Chapter 3 revealed a high reliability of methods based on differentiation between (sub)populations and showed that hard and soft selective sweeps are most reliably detected using a combination of metrics, including $\Delta\pi$, genotype distance distribution-based, and differentiation-based metrics, some of which have rarely been used in the context of selection scans in the past. The simulation results also showed that a recently developed beta diversity metric, applied for the first time to genetic data in Chapter 3, has high reliability for detecting selective sweeps, outcompeting the other metrics used. In line with the approach of using multiple metrics to infer signals of selection, the results of Chapter 3 highlight the importance of combining metrics that make use of different properties of the genome.

6.2 Discussion of the general relevance of results

The identified genes and gene functions under selection in Chapters 2, 4, and 5 can be considered as putative candidates of genetic factors contributing to rapid diversification in highly diversifying salmonids and potentially also in other taxonomic groups. They highlight the importance of neural processes, trophic behaviour, immunity, and muscle function in effective local adaptation. Neural functions in particular appeared in all top selection outlier sets of the above chapters, which matches the observation of many neural process genes in positive selection and gene regulation outlier sets of other rapidly diversifying taxa (Brawand et al. 2014; York et al. 2019; Liang et al. 2020). Similarly, feeding behaviour has also been implicated in adaptive divergence, for example in whitefish (Lundsgaard-Hansen et al. 2013) and Darwin's finches (Tebich et al. 2010). These observations have potential implications for answering the question of what genetic traits need to change in order for the phenotypes of ecomorphs to change. A large proportion of studies on rapid diversifications and adaptive radiations have focussed on visible morphological differences (e.g. body shape and jaw morphology), partly owing to the ease of measuring such traits as compared to physiological, neural, and behavioural traits. The results shown here imply that the underlying genes of such visible traits are not necessarily among the top outlier genes shared in rapid diversifications and adaptive divergence events. Moreover, genes involved in feeding behaviour are often capable of changing more easily visible morphological traits, such as jaw morphology, body size, and body shape, owing to the underlying phenotypic plasticity of these traits (Adams and Huntingford 2004; Alexander and Adams 2004; Wund et al. 2008; Pfennig et al. 2010). The observation of phenotypic plasticity in lab and common-garden experiments therefore likely has limited potential for making judgements about the role of genetic factors shaping ecomorphs in adaptive diversifications, but also underlines the role phenotypic plasticity can play in evolutionary adaptation.

The notion that gene regulation and changes in the regulatory regions of genes play a central role in adaptation is supported by our results. The importance of gene regulation in this process extends far beyond salmonids or even vertebrates, as in many model systems for adaptive diversification similar observations have been made (Mack and Nachman 2017; Singh et al. 2017; Franchini et al. 2019; Campana et al. 2020; Jacobs et al. 2020). This also shows that the predominance of conserved coding regions among outlier genes does not allow conclusions about the influence of natural selection on such genes. While gene expression is a plastic process, this does not mean that it has no influence on

seemingly static morphological, trophic, and neural traits that facilitate local adaptation of ecomorphs.

A function of similar generality for diversification in animals is immunity, which has seen repeated adaptations to local conditions (O'Connor et al. 2016; Svardal et al. 2020). Specifically for teleost fishes, immunity genes have been implicated in increasing speciation rates (Malmstrøm et al. 2016). This role of immunity is supported by the results in salmonids here and the observed prevalence of diversifying selection in immunity genes of highly diversifying salmonids. The evolution of immunity genes, however, is also influenced to a high degree by local environmental conditions and the presence of parasites and pathogens in a habitat (Early and Clark 2017; Lopez et al. 2019). Therefore, selection on immunity genes might be highly species-specific and dependent on the ecological conditions an organism faces.

Lastly, the observations made using the simulations of chromosomes in Chapter 3 emphasise how single metrics can fail to detect selective sweeps, particularly under some demographic scenarios. A combination of metrics or joint metrics of selection as well as more sophisticated approaches to infer the probability and significance of positive selection in a region, such as machine learning approaches, can prove more powerful than single metrics and are potentially more robust to the detrimental effects of complex or unknown demographic processes (Alachiotis and Pavlidis 2018; Kern and Schrider 2018; Schrider and Kern 2018).

6.3 Outlook and future directions

Several aspects of the study of rapid diversifications in salmonids would profit from additional research, experiments, and analyses. The functional significance of the outlier genes, GO categories, and regions found here in the context of intra- and interspecies analyses using whole-genome and transcriptome sequences remains to be tested in systematic experimental studies, such as gene expression assessments and targeted gene editing. Examining gene expression of outlier genes more closely seems particularly relevant given the lack of protein-coding changes in the outlier genes. A more in-depth analysis and more extensive sampling of populations of Arctic charr, whitefish, and also of other salmonid groups would allow conclusions about the allelic distribution in regions of selection, which could then be matched with the distribution of gene expression. Potentially missed allelic variants that could influence either the regulation or function of outlier genes, or both, could also be detected in this way. Moreover, the application of putatively more powerful methods to infer selection, some of which were presented here,

could improve, narrow down, or confirm the targets of selection in the above chapters (Kern and Schrider 2018; Schneider et al. 2020). Screening for genetic variants with strong effects on regulation or function of genes is a prerequisite for applying modern techniques of gene editing, such as CRISPR-Cas9 and gene knockouts (Ran et al. 2013; Doudna and Charpentier 2014; Fernandez i Marti and Dodd 2018; Chao et al. 2019). However, regulatory and functionally important variants can also be highly polygenic (Brawand et al. 2014; Laurentino et al. 2020), which was implied by the incomplete differentiation of variants obtained in Chapter 4, complicating the application of functional assays due to the high number of editing targets.

High-coverage sequencing of a subset of samples, especially in and around candidate genes, is necessary to make more conclusive judgements about structural variants that potentially affect gene function. Linkage maps of Arctic charr and other salmonids could be used to infer local recombination rates across the genome, which can pinpoint potential locations of inversions and other recombination-reducing structural variants (Kawakami et al. 2014). These locations could then be computationally screened for inversions (Mahmoud et al. 2019), especially in the vicinity of outlier regions showing signatures of selection, which could allow inferences about genetic variants reducing the local migration rate, thus facilitating reproductive isolation (Wellenreuther and Bernatchez 2018). This could be accompanied by computational prediction of gene function to narrow down targets for future gene editing studies (Tang and Thomas 2016; Lee et al. 2017).

It should also be noted that more extensive sampling of populations across the species range could improve judgements about the demographic history of diversifying Arctic charr and similarly for whitefish, which can also play a role in the prediction of signals of selection, as shown in Chapter 3. Furthermore, this would allow assessment of the commonness of adaptive introgression, as found in Chapter 5, at a more global level. Examining the generality of adaptive introgression of locally adapted genetic variants also has implications for the ecological diversification process in salmonids as a whole. Whether a pool of standing genetic variation exists in Arctic charr and other salmonid species, as shown for threespine sticklebacks (*Gasterosteus aculeatus*) (Feulner et al. 2013; Marques et al. 2018), remains to be tested. Future research will show the extent of generality of the results presented here.

Appendices

Appendix A: Figures

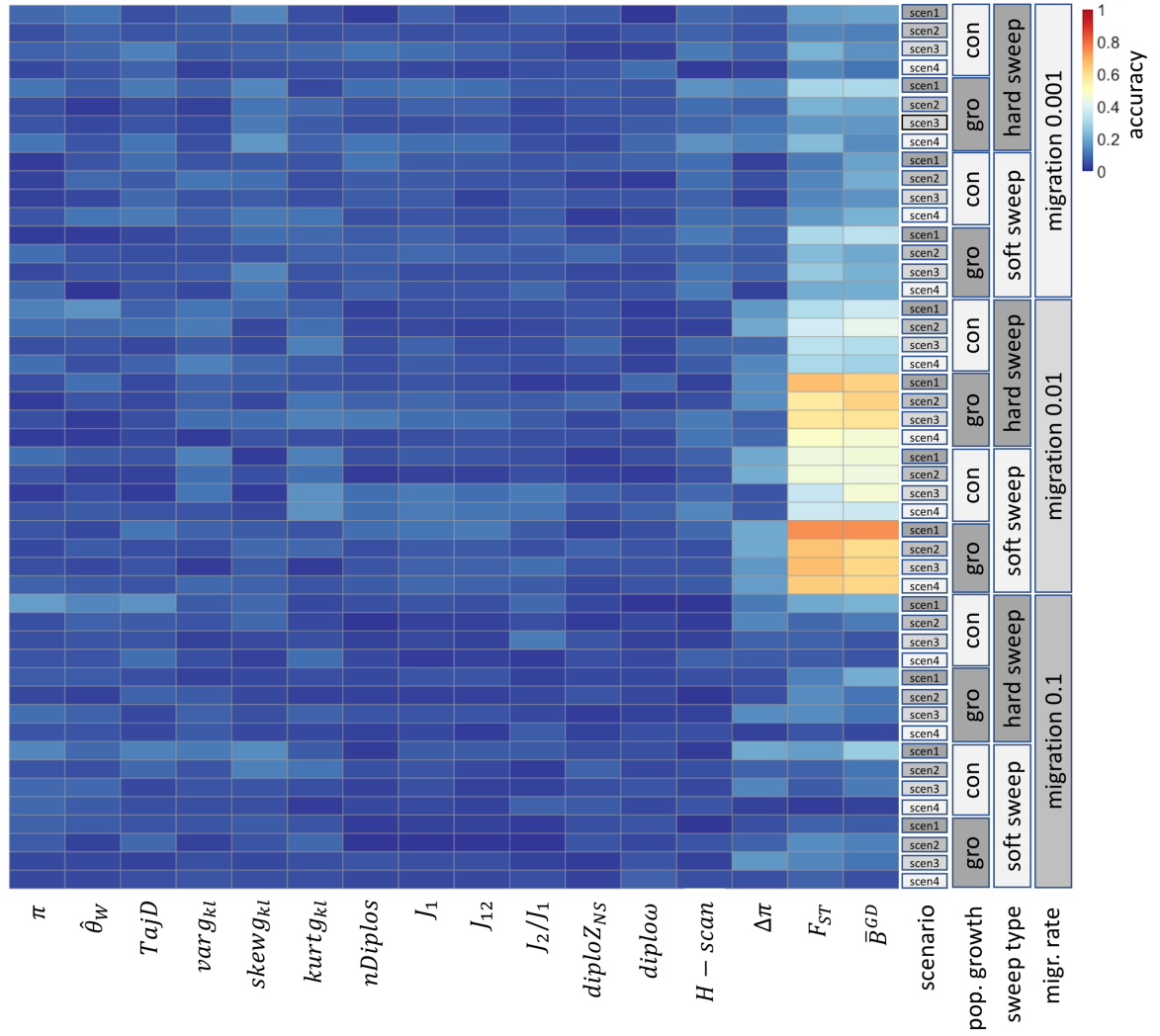


Figure S3.1 – A heat map of prediction accuracies of the location of selective sweeps for the low recombination rate region and equal population sizes. The colours in the heat map are based on accuracy values (see scale at the top right of the plot, with the highest accuracies, 1, in dark red and the lowest accuracies, 0, in dark blue), which are the proportions of 100 simulation replicates in which the correct window was identified as containing the selected mutation (1-in-1 window approach). For $\bar{B}^{GD}$, the chosen q value was 2. scen1 = scenario 1 (sympatric divergence); scen2 = scenario 2 (secondary contact scenario with low negative selection in population 3); scen3 = scenario 3 (secondary contact scenario with intermediate negative selection in population 3); scen4 = scenario 4 (secondary contact scenario with strong negative selection in population 3); gro = logistically increasing population size; con = constant population size; soft sweep = soft selective sweep scenario; hard sweep = hard selective sweep scenario; migration = migration rate between populations 2 and 3; pop. growth = population growth (gro or con); migr. rate = migration rate.

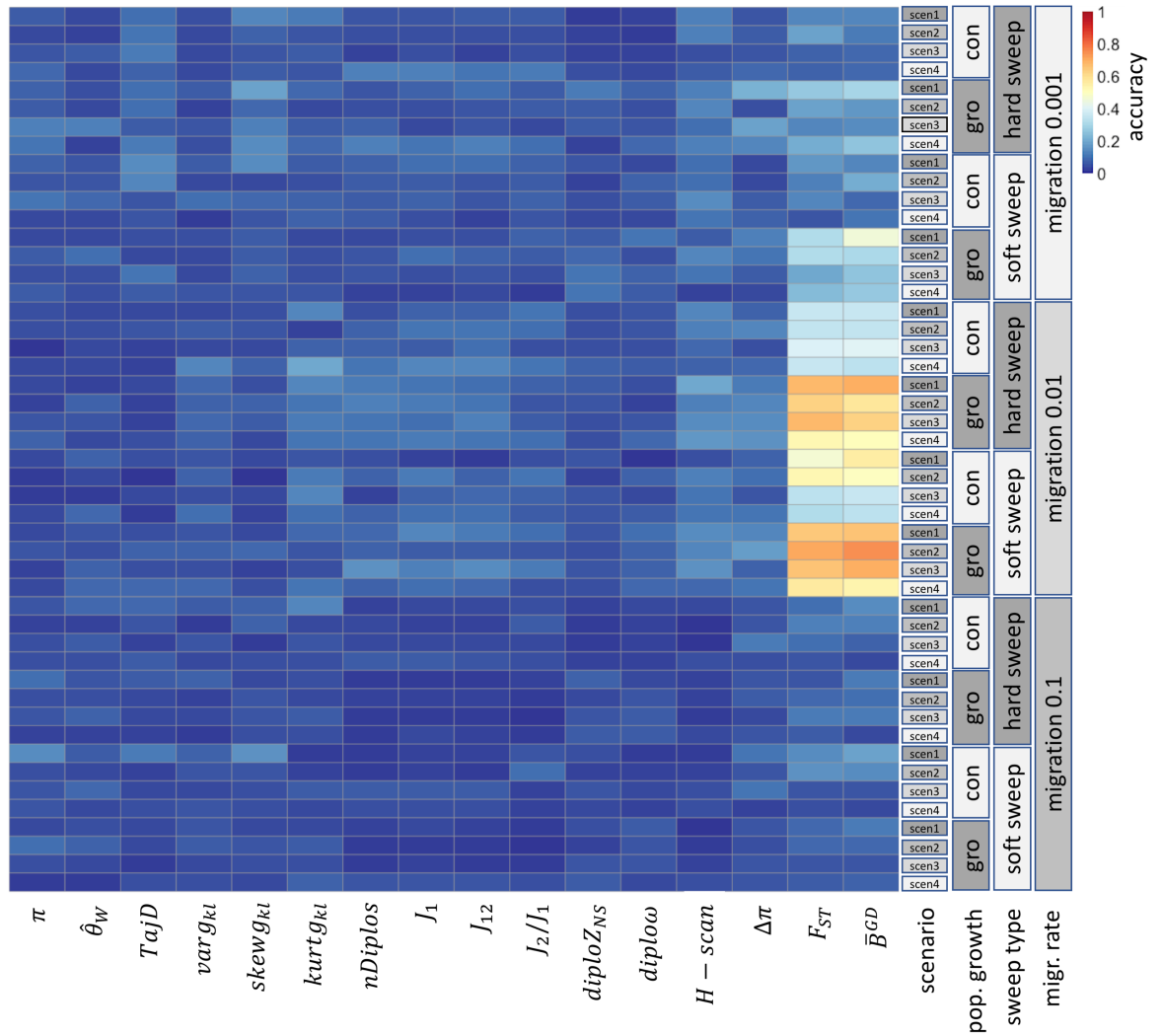


Figure S3.2 – A heat map of prediction accuracies of the location of selective sweeps for the low recombination rate region and unequal population sizes. The colours in the heat map are based on accuracy values (see scale at the top right of the plot, with the highest accuracies, 1, in dark red and the lowest accuracies, 0, in dark blue), which are the proportions of 100 simulation replicates in which the correct window was identified as containing the selected mutation (1-in-1 window approach). For $\bar{B}^{GD}$, the chosen q value was 2. scen1 = scenario 1 (sympatric divergence); scen2 = scenario 2 (secondary contact scenario with low negative selection in population 3); scen3 = scenario 3 (secondary contact scenario with intermediate negative selection in population 3); scen4 = scenario 4 (secondary contact scenario with strong negative selection in population 3); gro = logistically increasing population size; con = constant population size; soft sweep = soft selective sweep scenario; hard sweep = hard selective sweep scenario; migration = migration rate between populations 2 and 3; pop. growth = population growth (gro or con); migr. rate = migration rate.

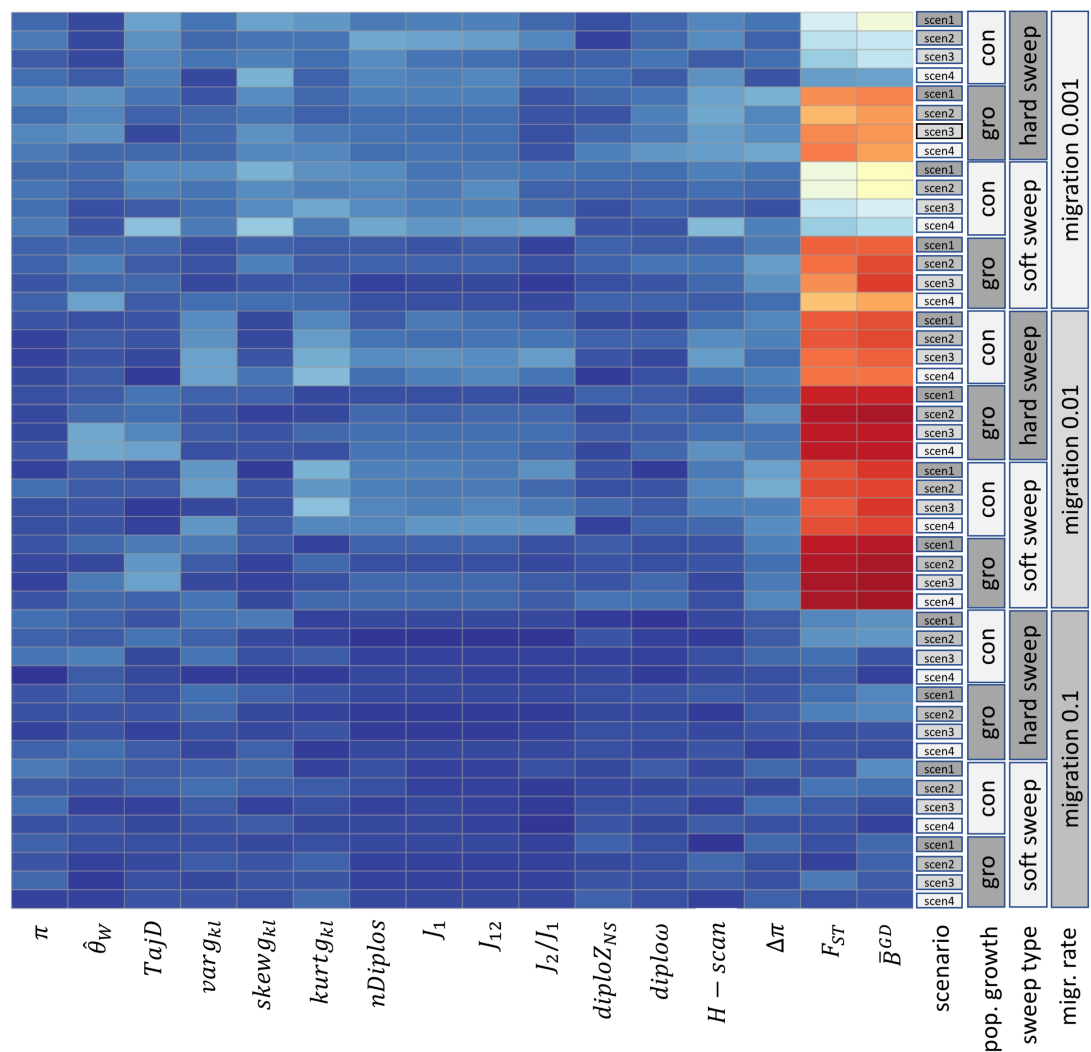


Figure S3.3 – A heat map of prediction accuracies of the location of selective sweeps for the standard recombination rate region and unequal population sizes. The colours in the heat map are based on accuracy values (see scale at the top right of the plot, with the highest accuracies, 1, in dark red and the lowest accuracies, 0, in dark blue), which are the proportions of 100 simulation replicates in which the correct window was identified as containing the selected mutation (1-in-1 window approach). For $\bar{B}^{GD}$, the chosen q value was 2. scen1 = scenario 1 (sympatric divergence); scen2 = scenario 2 (secondary contact scenario with low negative selection in population 3); scen3 = scenario 3 (secondary contact scenario with intermediate negative selection in population 3); scen4 = scenario 4 (secondary contact scenario with strong negative selection in population 3); gro = logistically increasing population size; con = constant population size; soft sweep = soft selective sweep scenario; hard sweep = hard selective sweep scenario; migration = migration rate between populations 2 and 3; pop. growth = population growth (gro or con); migr. rate = migration rate.

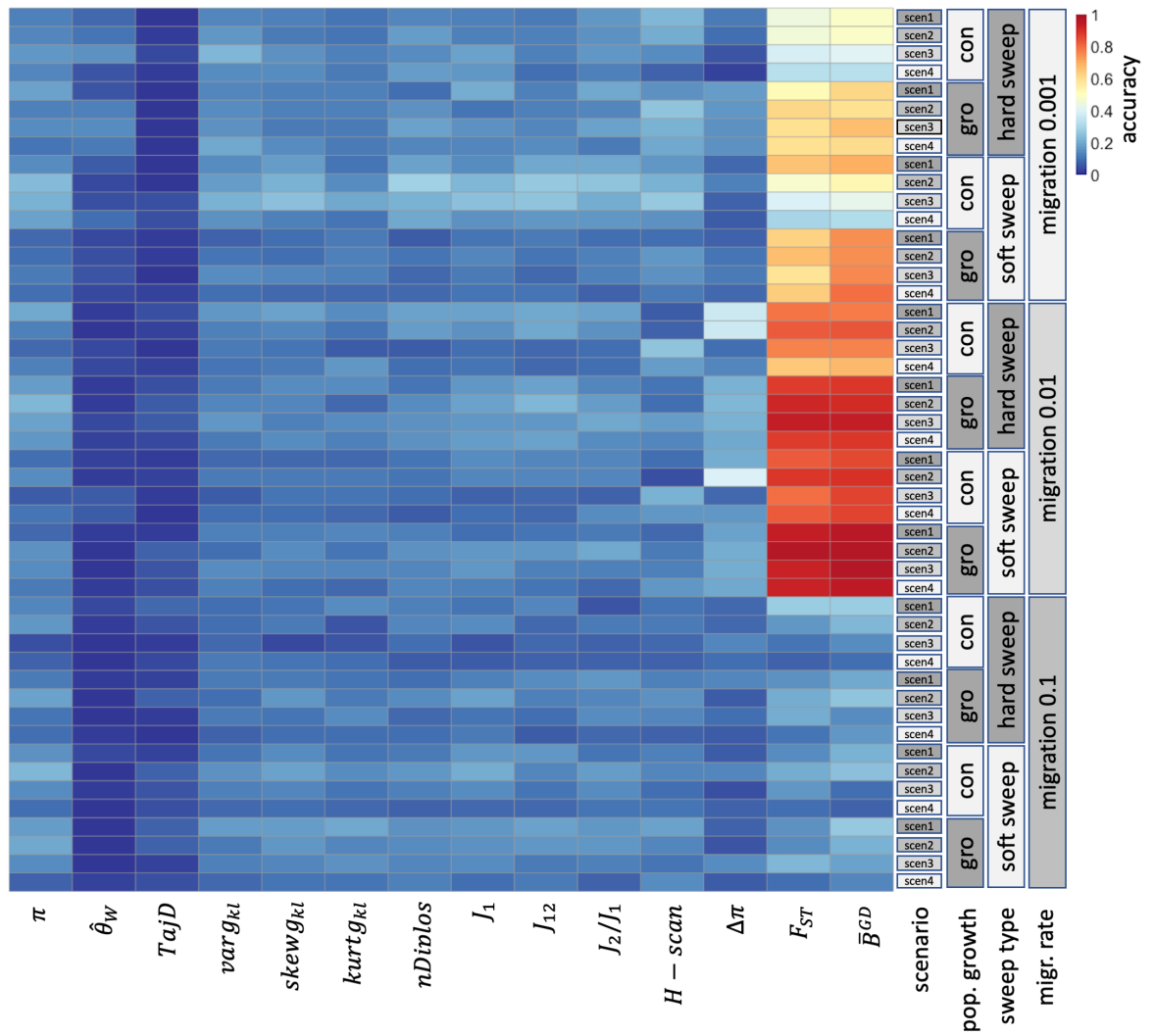


Figure S3.4 – A heat map of prediction accuracies of the location of selective sweeps for the standard recombination rate region and equal population sizes for the SNP window-based approach. The colours in the heat map are based on accuracy values (see scale at the top right of the plot, with the highest accuracies, 1, in dark red and the lowest accuracies, 0, in dark blue), which are the proportions of 100 simulation replicates in which the correct window was identified as containing the selected mutation (1-in-1 window approach). For $\bar{B}^{GD}$, the chosen q value was 2. scen1 = scenario 1 (sympatric divergence); scen2 = scenario 2 (secondary contact scenario with low negative selection in population 3); scen3 = scenario 3 (secondary contact scenario with intermediate negative selection in population 3); scen4 = scenario 4 (secondary contact scenario with strong negative selection in population 3); gro = logistically increasing population size; con = constant population size; soft sweep = soft selective sweep scenario; hard sweep = hard selective sweep scenario; migration = migration rate between populations 2 and 3; pop. growth = population growth (gro or con); migr. rate = migration rate.

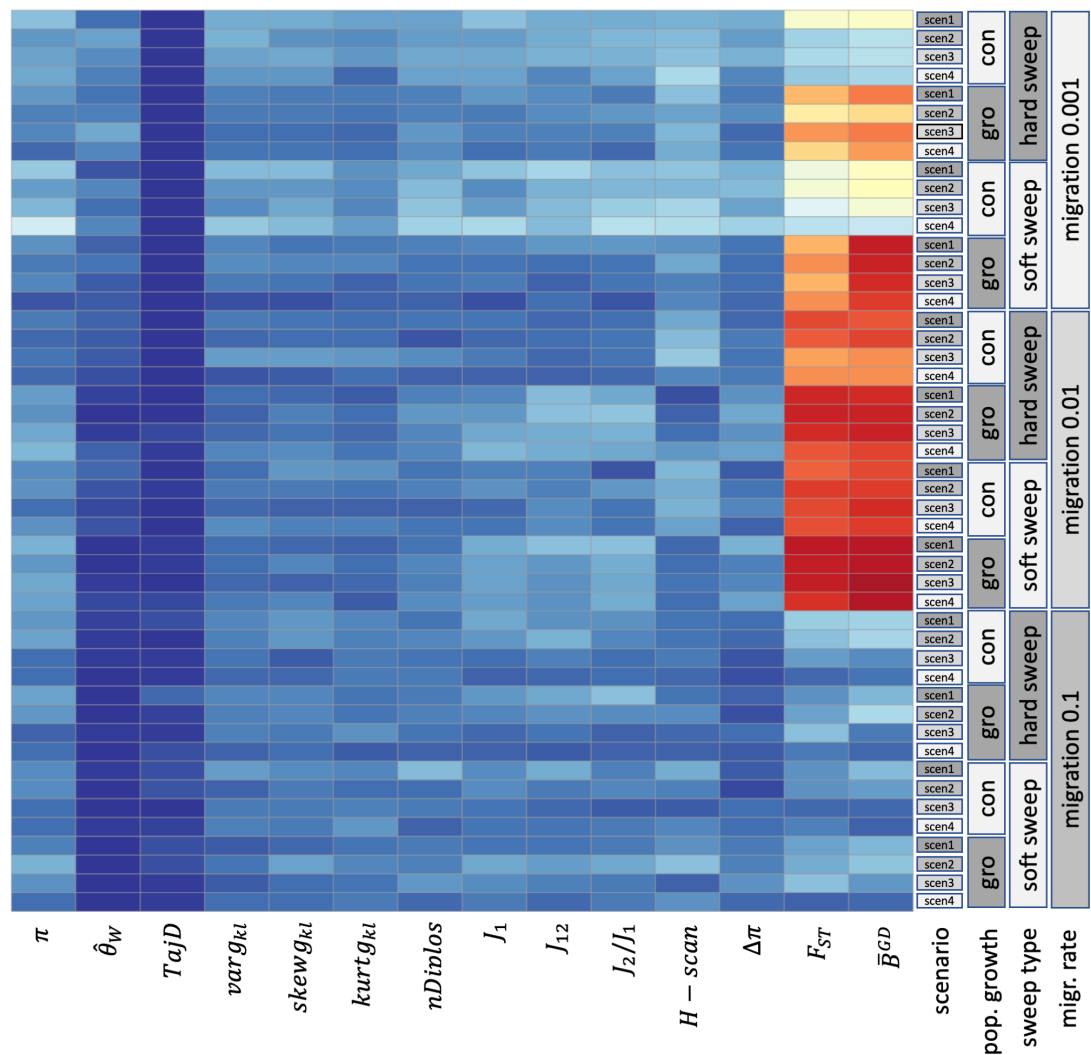


Figure S3.5 – A heat map of prediction accuracies of the location of selective sweeps for the standard recombination rate region and unequal population sizes for the SNP window-based approach. The colours in the heat map are based on accuracy values (see scale at the top right of the plot, with the highest accuracies, 1, in dark red and the lowest accuracies, 0, in dark blue), which are the proportions of 100 simulation replicates in which the correct window was identified as containing the selected mutation (1-in-1 window approach). For $\bar{B}^{GD}$, the chosen q value was 2. scen1 = scenario 1 (sympatric divergence); scen2 = scenario 2 (secondary contact scenario with low negative selection in population 3); scen3 = scenario 3 (secondary contact scenario with intermediate negative selection in population 3); scen4 = scenario 4 (secondary contact scenario with strong negative selection in population 3); gro = logistically increasing population size; con = constant population size; soft sweep = soft selective sweep scenario; hard sweep = hard selective sweep scenario; migration = migration rate between populations 2 and 3; pop. growth = population growth (gro or con); migr. rate = migration rate.

Figure S3.6 – A heat map of prediction accuracies of the location of selective sweeps for the low recombination rate region and equal population sizes for the SNP window-based approach. The colours in the heat map are based on accuracy values (see scale at the top right of the plot, with the highest accuracies, 1, in dark red and the lowest accuracies, 0, in dark blue), which are the proportions of 100 simulation replicates in which the correct window was identified as containing the selected mutation (1-in-1 window approach). For $\bar{B}^{GD}$, the chosen q value was 2. scen1 = scenario 1 (sympatric divergence); scen2 = scenario 2 (secondary contact scenario with low negative selection in population 3); scen3 = scenario 3 (secondary contact scenario with intermediate negative selection in population 3); scen4 = scenario 4 (secondary contact scenario with strong negative selection in population 3); gro = logistically increasing population size; con = constant population size; soft sweep = soft selective sweep scenario; hard sweep = hard selective sweep scenario; migration = migration rate between populations 2 and 3; pop. growth = population growth (gro or con); migr. rate = migration rate.

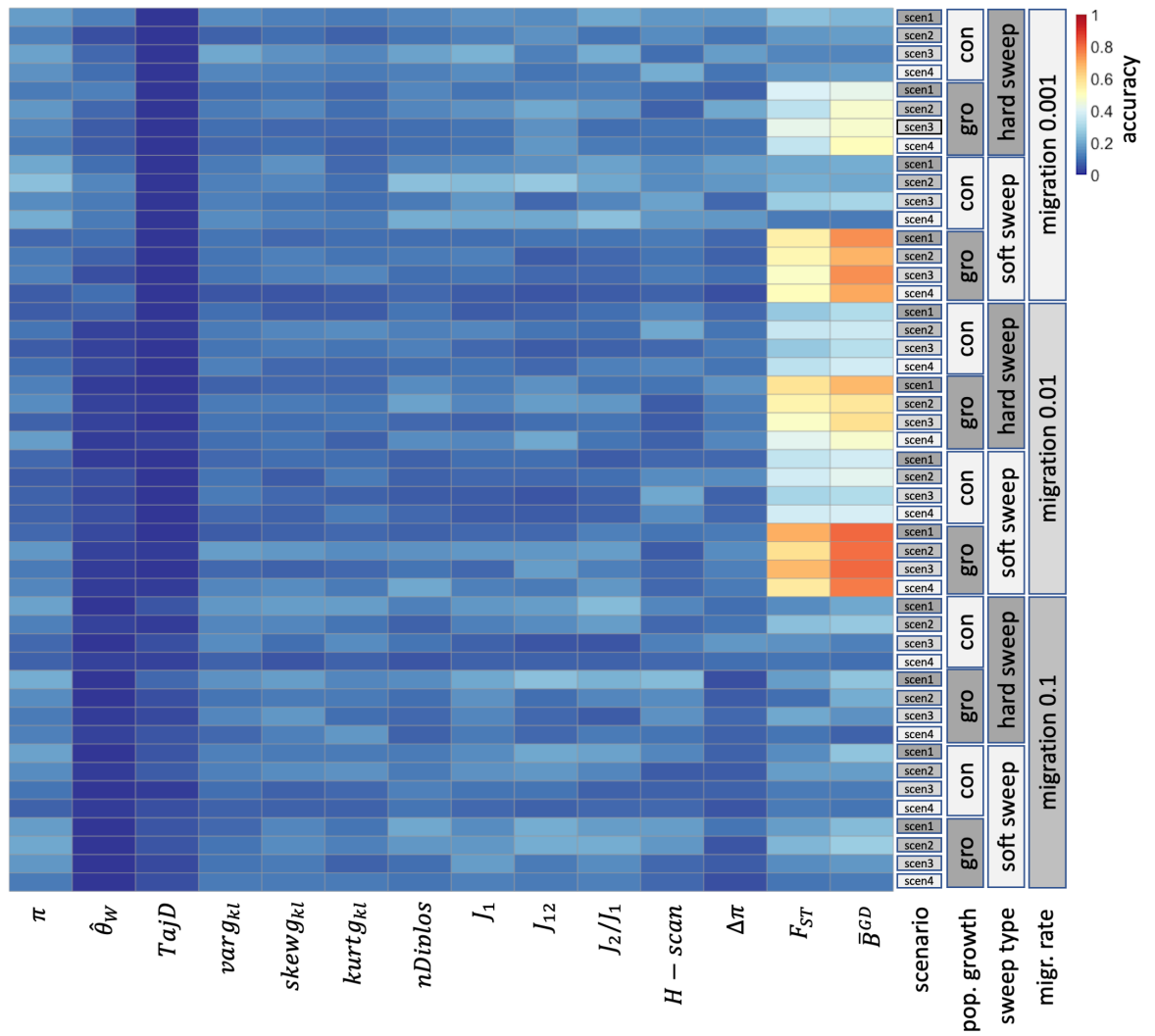


Figure S3.7 – A heat map of prediction accuracies of the location of selective sweeps for the low recombination rate region and unequal population sizes for the SNP window-based approach. The colours in the heat map are based on accuracy values (see scale at the top right of the plot, with the highest accuracies, 1, in dark red and the lowest accuracies, 0, in dark blue), which are the proportions of 100 simulation replicates in which the correct window was identified as containing the selected mutation (1-in-1 window approach). For $\bar{B}^{GD}$, the chosen q value was 2. scen1 = scenario 1 (sympatric divergence); scen2 = scenario 2 (secondary contact scenario with low negative selection in population 3); scen3 = scenario 3 (secondary contact scenario with intermediate negative selection in population 3); scen4 = scenario 4 (secondary contact scenario with strong negative selection in population 3); gro = logistically increasing population size; con = constant population size; soft sweep = soft selective sweep scenario; hard sweep = hard selective sweep scenario; migration = migration rate between populations 2 and 3; pop. growth = population growth (gro or con); migr. rate = migration rate.

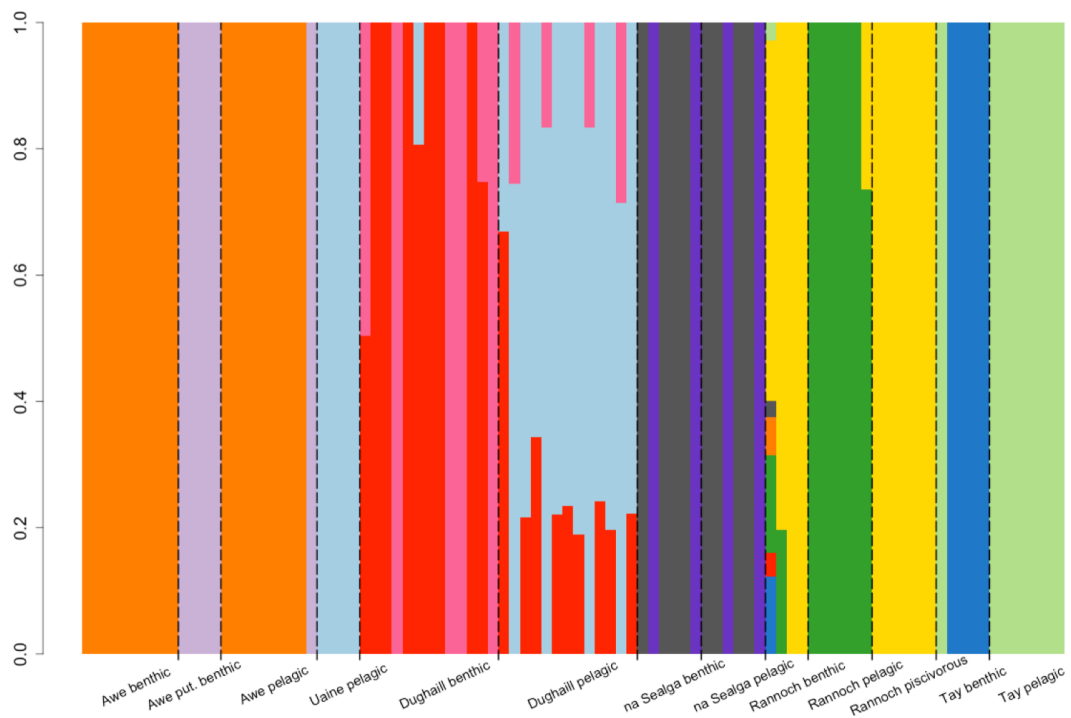
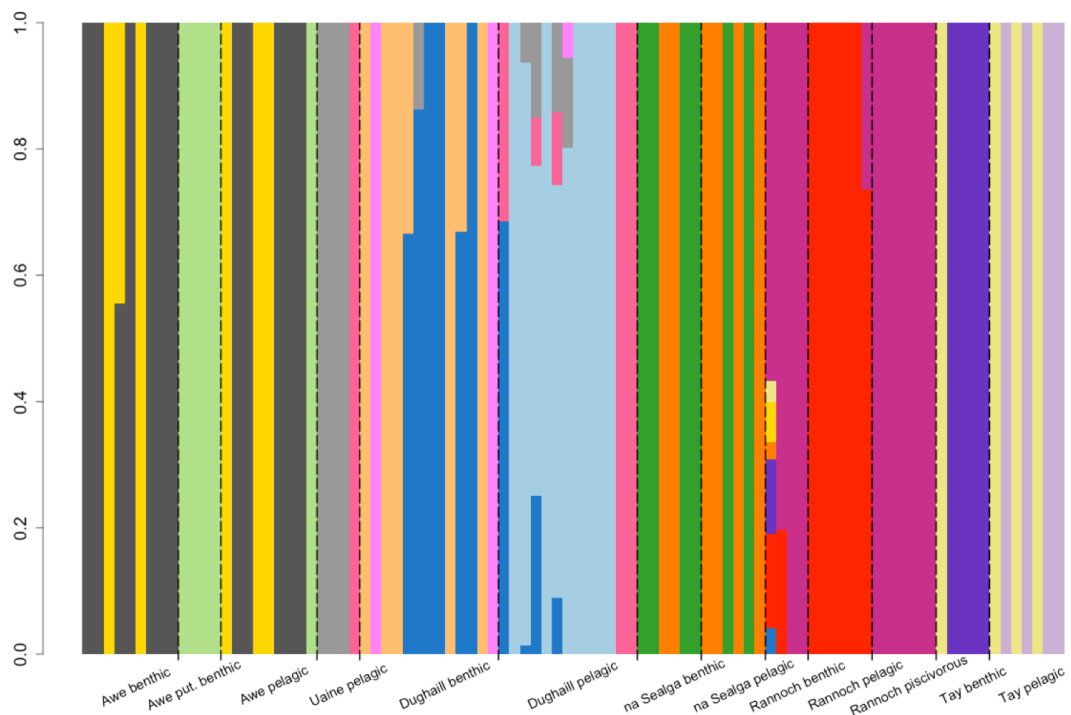
a)**b)**

Figure S4.1 – Individual genome-wide admixture patterns in Arctic charr across six lakes and up to three sympatric ecomorphs. The y-axis represents the proportion of the genome that clusters with a certain assumed ancestral population (K). The plots for an assumed ancestral number of populations of a) $K = 11$ and b) $K = 16$ are shown. Awe put. benthic = putatively benthic ecomorph of Loch Awe

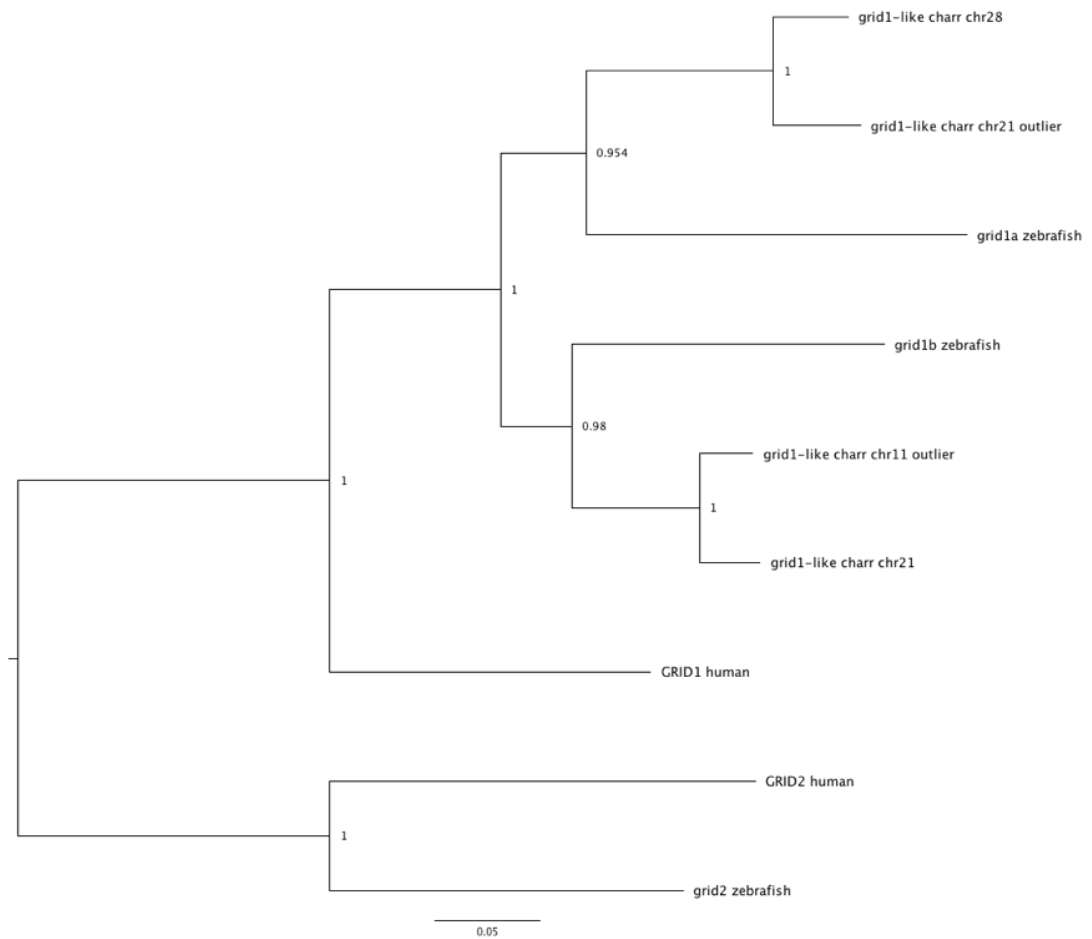


Figure S4.2 – Phylogenetic tree of *grid* genes in Arctic charr, human, and zebrafish. The values at the nodes are proportional bootstrap values calculated from 1000 bootstrap replications. The scale bar shows the average number of nucleotide changes per nucleotide site. outlier = version of the gene under selection; chr11 = chromosome 11; chr21 = chromosome 21

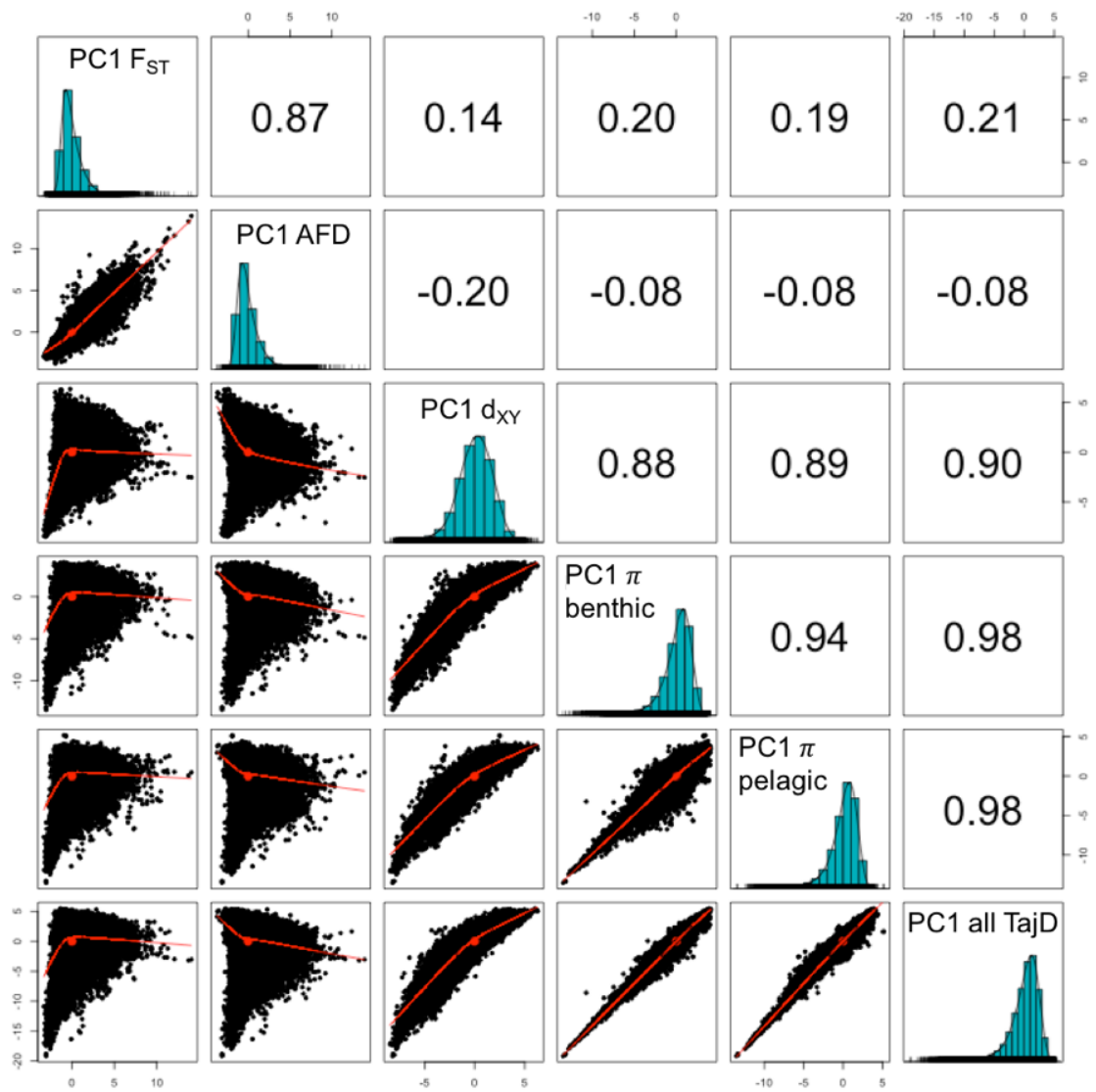


Figure S4.3 – Relationships and correlations between metric pairs. The bottom left triangle shows scatter plots and fitted spline curves. The top right triangle shows Pearson's correlation coefficients. The histograms in the diagonal squares are metric distributions.

TajD = *Tajima's D*

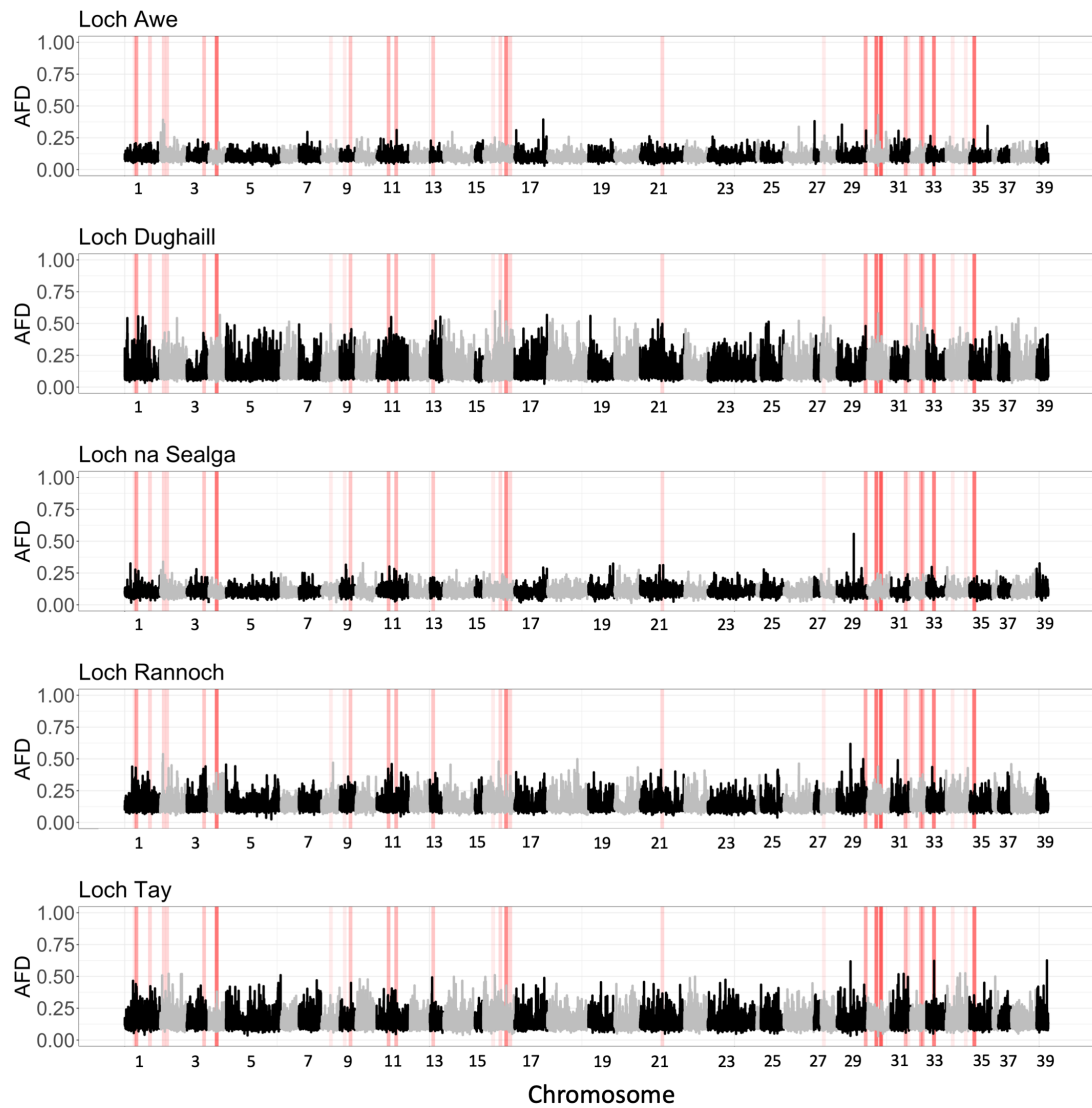


Figure S4.4 – Allele frequency difference (AFD) between benthic and pelagic ecomorphs. The landscape of AFD is shown separately for all five benthic-pelagic ecomorph pairs. The grey or black shadings indicate the 39 assembled chromosomes of Arctic charr. Chromosome ID numbers are shown below the AFD distributions. The red bars indicate shared outlier regions across AFD (top 5% quantile) and PCAngsd (FDR q value < 0.01) analyses across all five benthic-pelagic ecomorph pairs and *Tajima's D* (bottom 5% quantile) in the pelagic ecomorphs of the five main populations (approach 3 in 4.3.7). Unassembled portions of the genome, i.e. those with uncertain genomic location, were not included in these plots.

Appendix B: Tables

Table S2.1 – Over- and underrepresented terms from a GO enrichment analysis. Cor = *Coregonus*, Salv = *Salvelinus*, OGs = orthogroups.

GO category	<i>P</i> value	<i>P</i> value (Cor + Salv)	# in OGs under selection	# in reference OGs	# in OGs under selection (Cor + Salv)	# in reference OGs (Cor + Salv)
Overrepresented GO terms – relaxed selection						
PTW/PP1 phosphatase complex	2.55E-03	3.03E-03	2	0	2	0
acrosome reaction	2.55E-03	3.03E-03	2	0	2	0
integral component of membrane	3.96E-03	1.77E-02	21	205	17	163
transcription factor complex	4.59E-03	5.52E-02	11	78	8	68
regulation of transcription, DNA- templated	6.29E-03	3.12E-02	10	70	8	60
DNA-binding transcription factor activity	1.22E-02	5.98E-02	9	66	7	57
regulation of transcription by RNA polymerase II	1.35E-02	2.20E-01	4	15	2	14
lipid metabolic process	1.43E-02	8.76E-03	2	2	2	1
protein serine/threonine kinase activity	2.00E-02	1.48E-02	3	9	3	7
protein tyrosine phosphatase activity	2.30E-02	2.71E-02	2	3	2	3
phospholipid catabolic process	2.30E-02	2.71E-02	2	3	2	3
tyrosine metabolic process	3.34E-02	3.92E-02	2	4	2	4
oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen	3.34E-02	2.71E-02	2	4	2	3

serine family amino acid metabolic process	3.70E-02	1.96E-02	3	12	3	8
nucleus	4.42E-02	6.75E-02	10	98	9	84
Overrepresented GO terms – intensified selection						
protein deubiquitination	1.73E-03	1.33E-01	3	4	1	3
regulation of transcription, DNA-templated	2.17E-03	5.18E-03	10	81	8	70
thiol-dependent ubiquitinyl hydrolase activity	8.36E-03	1.02E-01	2	2	1	2
DNA binding	2.48E-02	1.68E-02	8	85	7	70
steroid hormone receptor activity	2.71E-02	2.29E-02	2	5	2	5
acetyl-CoA carboxylase activity	3.85E-02	3.52E-02	1	0	1	0
acetyl-CoA carboxylase complex	3.85E-02	3.52E-02	1	0	1	0
Lys48-specific deubiquitinase activity	3.85E-02	1.00E+00	1	0	0	0
Sin3 complex	3.85E-02	3.52E-02	1	0	1	0
histone H3-K4 demethylation	3.85E-02	3.52E-02	1	0	1	0
retinoic acid biosynthetic process	3.85E-02	3.52E-02	1	0	1	0
NADP-retinol dehydrogenase activity	3.85E-02	3.52E-02	1	0	1	0
catalase activity	3.85E-02	1.00E+00	1	0	0	0
obsolete peroxidase reaction	3.85E-02	1.00E+00	1	0	0	0
response to oxidative stress	3.85E-02	1.00E+00	1	0	0	0
methane metabolic process	3.85E-02	1.00E+00	1	0	0	0
hydrogen peroxide catabolic process	3.85E-02	1.00E+00	1	0	0	0
palmitoyltransferase activity	3.85E-02	3.52E-02	1	0	1	0

beta1-adrenergic receptor activity	3.85E-02	1.00E+00	1	0	0	0
adenylate cyclase-activating G protein-coupled receptor signaling pathway	3.85E-02	1.00E+00	1	0	0	0
positive regulation of heart contraction	3.85E-02	1.00E+00	1	0	0	0
adrenergic receptor signaling pathway	3.85E-02	1.00E+00	1	0	0	0
adrenal gland development	3.85E-02	3.52E-02	1	0	1	0
corticotropin-releasing hormone binding	3.85E-02	1.00E+00	1	0	0	0
apoptotic signaling pathway	3.85E-02	3.52E-02	1	0	1	0
regulation of NIK/NF-kappaB signaling	3.85E-02	3.52E-02	1	0	1	0
iron-sulfur cluster binding	3.85E-02	3.52E-02	1	0	1	0
actin filament severing	3.85E-02	3.52E-02	1	0	1	0
dystrophin-associated glycoprotein complex	3.85E-02	3.52E-02	1	0	1	0
hexosaminidase activity	3.85E-02	1.00E+00	1	0	0	0
hyaluronic acid binding	3.85E-02	3.52E-02	1	0	1	0
ribosomal large subunit biogenesis	3.85E-02	3.52E-02	1	0	1	0
Overrepresented GO terms – diversifying selection						
transmembrane transporter activity	2.28E-02	1.19E-02	4	23	4	18
transmembrane transport	2.17E-02	1.25E-02	6	48	6	41
phospholipid metabolic process	4.11E-02	4.20E-02	1	0	1	0
regulation of Wnt signaling pathway	4.11E-02	4.20E-02	1	0	1	0
lipid phosphatase activity	4.11E-02	4.20E-02	1	0	1	0
proteolysis	3.20E-02	2.18E-02	6	53	6	47

acetyl-CoA carboxylase activity	4.11E-02	4.20E-02	1	0	1	0
pyruvate metabolic process	1.67E-03	1.75E-03	2	0	2	0
acetyl-CoA carboxylase complex	4.11E-02	4.20E-02	1	0	1	0
NAD-dependent histone deacetylase activity (H3-K14 specific)	4.11E-02	4.20E-02	1	0	1	0
histone H3 deacetylation	4.11E-02	4.20E-02	1	0	1	0
intracellular signal transduction	1.51E-02	4.27E-02	4	20	3	16
transforming growth factor beta receptor signaling pathway	4.11E-02	4.20E-02	1	0	1	0
serine-type endopeptidase activity	5.27E-03	4.56E-03	4	14	4	13
cytoskeleton	4.98E-02	4.14E-02	2	7	2	6
septin complex	4.11E-02	4.20E-02	1	0	1	0
complement activation	4.11E-02	4.20E-02	1	0	1	0
nociceptin receptor activity	4.11E-02	1.00E+00	1	0	0	0
calcium-mediated signaling	4.11E-02	4.20E-02	1	0	1	0
glycerolipid metabolic process	3.98E-02	4.14E-02	2	6	2	6
kinesin complex	4.11E-02	4.20E-02	1	0	1	0
kinesin binding	4.11E-02	4.20E-02	1	0	1	0
lipid catabolic process	4.88E-03	5.10E-03	2	1	2	1
oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen	2.25E-02	1.93E-01	2	4	1	4
L-ascorbic acid binding	1.54E-02	1.58E-01	2	3	1	3
lipoprotein lipase activity	1.67E-03	1.75E-03	2	0	2	0

Lys48-specific deubiquitinase activity	4.11E-02	1.00E+00	1	0	0	0
synapse	4.11E-02	4.20E-02	1	0	1	0
Sin3 complex	4.11E-02	4.20E-02	1	0	1	0
sulfate assimilation	4.11E-02	4.20E-02	1	0	1	0
adenylylsulfate kinase activity	4.11E-02	4.20E-02	1	0	1	0
sulfate adenylyltransferase (ATP) activity	4.11E-02	4.20E-02	1	0	1	0
glycogen metabolic process	4.11E-02	4.20E-02	1	0	1	0
antigen processing and presentation of endogenous peptide antigen via MHC class I	4.11E-02	4.20E-02	1	0	1	0
melatonin receptor activity	4.11E-02	1.00E+00	1	0	0	0
3-hydroxyisobutyrate dehydrogenase activity	4.11E-02	4.20E-02	1	0	1	0
NAD binding	1.54E-02	1.61E-02	2	3	2	3
urea transmembrane transporter activity	4.11E-02	4.20E-02	1	0	1	0
urea transmembrane transport	4.11E-02	4.20E-02	1	0	1	0
5-methylcytosine catabolic process	4.11E-02	4.20E-02	1	0	1	0
methylcytosine dioxygenase activity	4.11E-02	4.20E-02	1	0	1	0
DNA demethylation	4.11E-02	4.20E-02	1	0	1	0
pyrroline-5-carboxylate reductase activity	4.11E-02	4.20E-02	1	0	1	0
proline biosynthetic process	4.11E-02	4.20E-02	1	0	1	0
malate dehydrogenase (decarboxylating) (NAD+) activity	4.11E-02	4.20E-02	1	0	1	0
malate metabolic process	4.11E-02	4.20E-02	1	0	1	0

proteasome regulatory particle assembly	4.11E-02	4.20E-02	1	0	1	0
beta1-adrenergic receptor activity	4.11E-02	1.00E+00	1	0	0	0
adenylate cyclase-activating G protein-coupled receptor signaling pathway	4.11E-02	1.00E+00	1	0	0	0
positive regulation of heart contraction	4.11E-02	1.00E+00	1	0	0	0
adrenergic receptor signaling pathway	4.11E-02	1.00E+00	1	0	0	0
small ribosomal subunit	4.11E-02	4.20E-02	1	0	1	0
10-formyltetrahydrofolate catabolic process	4.11E-02	4.20E-02	1	0	1	0
formyltetrahydrofolate dehydrogenase activity	4.11E-02	4.20E-02	1	0	1	0
hydroxymethyl-, formyl- and related transferase activity	4.11E-02	4.20E-02	1	0	1	0
regulation of neuroinflammatory response	4.11E-02	1.00E+00	1	0	0	0
hexosaminidase activity	4.11E-02	1.00E+00	1	0	0	0
junctional membrane complex	4.11E-02	4.20E-02	1	0	1	0
Underrepresented GO terms – diversifying selection						
DNA binding	1.89E-02	3.46E-02	0	93	0	77

Table S2.2 – Orthogroups (OGs) with signals of relaxed and diversifying selection or intensified and diversifying selection. k values are selection parameter values (exponent of dN/dS ratio, i.e. k in ω^k). FDR values are from Benjamini-Hochberg FDR correction. Orthogroups shown are present in both *Coregonus* and *Salvelinus*. For aBSREL, P values after correction for multiple testing (FDR < 0.10) are shown.

OG ID	Gene description	GO terms (obtained with Blast2GO)	Additional information (manually derived from UniProt/Swiss-Prot)	FDR (RELAX)	p (aBSREL)	k
Relaxed & diversifying selection						
OG0001544	adaptor-related protein complex 1, mu 2 subunit (ap1m2)	intracellular protein transport; vesicle-mediated transport; clathrin adaptor complex	antigen processing and presentation of exogenous peptide antigen via MHC class II, regulation of defence response to virus by virus	0.000	1.019E-05	0.04471
OG0002461	protein kinase C epsilon type-like (LOC106589411)	protein serine/threonine kinase activity; ATP binding; protein phosphorylation; serine family amino acid metabolic process; intracellular signal transduction	cell adhesion, motility, and cell cycle, neuron growth, ion channel regulation, immune response, signal transduction	6.825E-08	6.478E-06	0.07476
OG0003986	solute carrier family 6 (neurotransmitter transporter), member 8 (slc6a8)	neurotransmitter:sodium symporter activity; cation transport; neurotransmitter transport; integral component of membrane	muscle contraction, required for the uptake of creatine in muscles and brain	7.970E-05	3.046E-12	0.05603
OG0000840	probable histone deacetylase 1-B (hda1b)	nucleus; NAD-dependent histone deacetylase activity (H3-K14 specific); metal ion binding; histone H3 deacetylation	transcriptional regulation, cell cycle progression, developmental events	0.0002432	1.952E-06	2.398E-12
OG0005675	E3 ubiquitin-protein ligase RNF128-like (LOC110537269), transcript variant X1	–	interleukin regulation, T-cell control, functions in the patterning of the dorsal ectoderm, sensitises ectoderm to respond to neural-inducing signals	0.0002432	1.282E-08	0.5357
OG0004474	endoplasmic reticulum	endoplasmic reticulum membrane; proteolysis; regulation of blood pressure; zinc ion binding; antigen	regulation of innate immune response, response to bacterium,	0.01917	2.008E-05	0.1651

	aminopeptidase 1-like (LOC106570844)	processing and presentation of endogenous peptide antigen via MHC class I; metalloaminopeptidase activity	angiogenesis, regulation of blood pressure, fat cell differentiation			
OG0000651	tyrosine-protein phosphatase non-receptor type 6-like (LOC106576571), transcript variant X3	protein tyrosine phosphatase activity; protein dephosphorylation; tyrosine metabolic process	haematopoiesis, various immune functions	0.03186	1.547E-06	0.3844
OG0002139	calciopressin-2-like (LOC109900784), transcript variant X1	nucleic acid binding; calcium-mediated signaling	regulation of calcineurin-NFAT signalling cascade, calcium-mediated signalling, potential role in central nervous development	0.06629	2.753E-11	0.6235
OG0014882	junctionophilin-2 (LOC111976619)	junctional membrane complex	calcium ion homeostasis, regulation of cardiac muscle tissue development	0.08725	3.460E-09	0.6595
Intensified & diversifying selection						
OG0000798	acetyl-CoA carboxylase beta (acacb), transcript variant X3	acetyl-CoA carboxylase activity; ATP binding; pyruvate metabolic process; fatty acid biosynthetic process; acetyl-CoA carboxylase complex; metal ion binding	regulation of cholesterol biosynthetic process, acetyl-CoA metabolic process	0.000	0.000	4.683
OG0011982	synapsin-3 (LOC111950806)	ATP binding; neurotransmitter secretion; synaptic vesicle	neurotransmitter secretion, regulation of synaptic transmission, GABAergic	0.0001606	0.000	8.330
OG0004141	clone Contig3164 cathepsin K precursor	proteolysis; cysteine-type peptidase activity	osteoclastic bone resorption, bone remodelling, collagen metabolic process	0.0002432	3.664E-14	7.769
OG0013215	lysine-specific demethylase 4B-like (LOC106574093), transcript variant X3	histone H3-K36 demethylation; histone H3-K9 demethylation (manually derived from UniProt/Swiss-Prot)	histone H3-K36 demethylation, histone H3-K9 demethylation, negative regulation of transcription	0.003614	3.970E-12	10.28

OG0003822	sterile alpha motif domain-containing protein 9-like (LOC109881303), transcript variant X2	protein binding	endosomal vesicle fusion	0.006278	9.533E-06	1.603
OG0007378	alpha-2,8-sialyltransferase 8F-like (LOC106588277), transcript variant X2	protein glycosylation; sialyltransferase activity	carbohydrate biosynthetic process, glycolipid metabolic process, ganglioside biosynthetic process, (modulation of cell signal transduction, maybe involved in function of nervous system)	0.008768	2.090E-11 9.549E-08	1.9625
OG0003820	paired amphipathic helix protein Sin3a-like (LOC106587652), transcript variant X3	transcription corepressor activity; transcription factor complex; negative regulation of transcription	negative regulation of transcription, regulation of lipid metabolic process, regulation of circadian rhythms	0.01812	5.292E-09	1.450
OG0004629	furin-1-like (LOC106587394), transcript variant X2	serine-type endopeptidase activity; proteolysis	involved in microbial infection, regulation of endopeptidase activity, TGF-beta-1 activation	0.01896	0.000	19.99
OG0010786	serine/threonine-protein kinase 6 (stk6)	protein serine/threonine kinase activity; ATP binding; protein phosphorylation; serine family amino acid metabolic process	activation of protein kinase activity, regulation of transcription	0.01972	0.000	6.183
OG0012826	beta-1,4-galactosyltransferase 6-like (LOC106560385), transcript variant X1	carbohydrate metabolic process; transferase activity	neuronal maturation, axonal and myelin formation, lactosylceramide biosynthetic process, carbohydrate metabolic process	0.03015	1.221E-12	6.768
OG0010420	torsin-1A-interacting protein 2 (LOC111975742)	endoplasmic reticulum organization; positive regulation of ATPase activity; protein localization to nuclear envelope (manually derived from UniProt/Swiss-Prot)	endoplasmic reticulum integrity, positive regulation of ATPase activity	0.03800	0.000	1.870

OG0011057	receptor-interacting serine/threonine- protein kinase 4-like (LOC106567956)	protein kinase activity; protein binding; ATP binding; protein phosphorylation	positive regulation of NF-kappaB transcription factor activity	0.08386	5.012E-06	3.356
-----------	--	--	---	---------	-----------	-------

Table S3.1 – Overview of varied parameters in simulated scenarios – all scenario and parameter combinations were simulated (144 different scenarios in total; 100 replicates per scenario). p2 is the population with a selective sweep; p3 is the population without a selective sweep. Every simulation includes a 2,000,025 bp low recombination and a 2,000,025 bp standard recombination region

Scenario type (incl. relative fitness of carriers of mutation in p3)

Sympatric divergence (carriers of mutation in p3 have relative fitness = $(1 / \text{relative fitness in p2})$)

Secondary contact (carriers of mutation in p3 have relative fitness = $(1 / \text{relative fitness in p2})$)

Secondary contact (carriers of mutation in p3 have relative fitness = $(1 / \text{relative fitness in p2}) - 0.1$)

Secondary contact (carriers of mutation in p3 have relative fitness = $(1 / \text{relative fitness in p2}) - 0.5$)

Sweep type

Hard sweep

Soft sweep

No sweep (neutral simulation)

Migration rate in sympatric phase

0.1

0.01

0.001

Population growth

Logistic population growth (from 500 to 2500)

No population growth (constant size of 500)

Equal or unequal population size

Equal population size

Population size larger in p3 (2000 in case of constant size; from 2000 to 10,000 in case of logistic growth)

Table S4.1 – *Grid* genes of Arctic charr, human, and zebrafish

Accession no.	Gene symbol	Gene name	Species	Inferred gene version (for charr)
XM_023970525.1	<i>grid1</i>	Glutamate receptor ionotropic, delta-1	<i>Salvelinus alpinus</i>	<i>grid1a</i>
XM_024008411.1	<i>grid1</i>	Glutamate receptor ionotropic, delta-1	<i>Salvelinus alpinus</i>	<i>grid1b</i>
XM_023993063.1	<i>grid1</i>	Glutamate receptor ionotropic, delta-1	<i>Salvelinus alpinus</i>	<i>grid1b</i>
XM_024008802.1	<i>grid1</i>	Glutamate receptor ionotropic, delta-1	<i>Salvelinus alpinus</i>	<i>grid1a</i>
NM_017551.3	<i>GRID1</i>	Glutamate receptor ionotropic, delta-1	<i>Homo sapiens</i>	
NM_001510.4	<i>GRID2</i>	Glutamate receptor ionotropic, delta-2	<i>Homo sapiens</i>	
XM_021467422.1	<i>grid1a</i>	Glutamate receptor ionotropic, delta-1a	<i>Danio rerio</i>	
XM_009306536.3	<i>grid1b</i>	Glutamate receptor ionotropic, delta-1b	<i>Danio rerio</i>	
AB154207.1	<i>grid2</i>	Glutamate receptor ionotropic, delta-2	<i>Danio rerio</i>	

Table S4.2 – Sampling and coverage information

Population/Ecomorph	Number of individuals	Mean no. of reads (Million)	Mean genome-wide coverage	Mean coverage in covered regions (≥1X)
Awe	22	79.09	2.54	6.27
Awe-benthic	9	81.13	2.54	6.47
Awe-pelagic	9	80.33	2.58	6.45
Awe-piscivorous	4	71.7	2.48	5.4
Dùghaill	26	156.58	4.95	10.26
Dùghaill-benthic	13	155.05	4.7	9.95
Dùghaill-pelagic	13	158.1	5.2	10.58
na Sealga	12	74.97	2.47	5.92
na Sealga-benthic	6	68.78	2.49	6.56
na Sealga-pelagic	6	81.17	2.45	5.28
Rannoch	16	114.77	3.67	9.04
Rannoch-benthic	4	113.92	3.7	8.74
Rannoch-pelagic	6	113.96	3.62	9
Rannoch-piscivorous	6	116.16	3.69	9.29
Tay	12	75.96	2.46	5.84
Tay-benthic	5	73.65	2.27	5.79
Tay-pelagic	7	77.62	2.59	5.87
Uaine	4	136.8	4.32	8.77
Total	92	108.76	3.48	7.89

Table S4.3 – NGSadmix log-likelihoods. The negative log-likelihoods for K values between 2 and 16 and the associated number of iterations needed to reach convergence are shown.

K	log-likelihood	N iterations
2	-952077786.9	447
3	-931169052.1	650
4	-907755957.3	900
5	-882347494.4	700
6	-866697350.8	1700
7	-857170124.2	500
8	-851831749.6	2000
9	-841402070	2000
10	-835081374.7	1300
11	-826530028.7	1050
12	-820044444.5	1250
13	-814517240	2000
14	-809033890	2000
15	-803162290.6	2000
16	-799178303.4	2000

Table S4.4 – Four-taxon test results for lochs Rannoch and Tay

D	Z	P value	nABBA	nBABA	n Blocks	H1	H2	H3	H4	Conclusion
-0.136	-94.077	0	198745	261405	13225	Rannoch_benthic	Rannoch_pelagic	Rannoch_piscivorous	naSealga_benthic	Ran ben closer to Ran pisc than to Ran pel
-0.019	-13.876	0	207046	215007	13063	Rannoch_benthic	Rannoch_pelagic	Tay_benthic	naSealga_benthic	Ran ben closer to Tay ben than to Ran pel
0.002	1.351	0.177	209171	208435	13116	Rannoch_benthic	Rannoch_pelagic	Tay_pelagic	naSealga_benthic	Tay pel equally close to Ran ben as to Ran pel
0.005	4.688	0.000	198745	196760	13198	Rannoch_benthic	Rannoch_piscivorous	Rannoch_pelagic	naSealga_benthic	Ran pisc potentially closer to Ran pel than to Ran ben
0.000	0.324	0.746	190427	190297	12986	Rannoch_benthic	Rannoch_piscivorous	Tay_benthic	naSealga_benthic	Tay ben equally close to Ran ben as to Ran pisc
0.001	0.679	0.497	188633	188369	13057	Rannoch_benthic	Rannoch_piscivorous	Tay_pelagic	naSealga_benthic	Tay pel equally close to Ran ben as to Ran pisc
-0.072	-46.927	0	207046	239017	13168	Rannoch_benthic	Tay_benthic	Rannoch_pelagic	naSealga_benthic	Ran ben closer to Ran pel than to Tay ben
-0.188	-121.049	0	190427	278761	13127	Rannoch_benthic	Tay_benthic	Rannoch_piscivorous	naSealga_benthic	Ran ben closer to Ran pisc than to Tay ben
0.086	58.051	0	237793	200277	13067	Rannoch_benthic	Tay_benthic	Tay_pelagic	naSealga_benthic	Tay ben closer to Tay pel than to Ran ben
-0.077	-49.221	0	209171	243910	13237	Rannoch_benthic	Tay_pelagic	Rannoch_pelagic	naSealga_benthic	Ran ben closer to Ran pel than to Tay pel
-0.209	-130.050	0	188633	288596	13221	Rannoch_benthic	Tay_pelagic	Rannoch_piscivorous	naSealga_benthic	Ran ben closer to Ran pisc than to Tay pel
0.058	36.589	0	237793	211868	13086	Rannoch_benthic	Tay_pelagic	Tay_benthic	naSealga_benthic	Tay pel closer to Tay ben than to Ran ben
0.141	98.223	0	261405	196760	13209	Rannoch_pelagic	Rannoch_piscivorous	Rannoch_benthic	naSealga_benthic	Ran pisc closer to Ran ben than to Ran pel
0.019	13.995	0	215039	206970	13103	Rannoch_pelagic	Rannoch_piscivorous	Tay_benthic	naSealga_benthic	Ran pisc closer to Tay ben than to Ran pel
-0.001	-0.867	0.386	208685	209160	13162	Rannoch_pelagic	Rannoch_piscivorous	Tay_pelagic	naSealga_benthic	Tay pel equally close to Ran pel as to Ran pisc
-0.053	-35.216	0	215007	239017	13161	Rannoch_pelagic	Tay_benthic	Rannoch_benthic	naSealga_benthic	Ran pel closer to Ran ben than to Tay ben
-0.057	-38.116	0	215039	240977	13223	Rannoch_pelagic	Tay_benthic	Rannoch_piscivorous	naSealga_benthic	Ran pel closer to Ran pisc than to Tay ben
0.083	59.299	0	239425	202620	13171	Rannoch_pelagic	Tay_benthic	Tay_pelagic	naSealga_benthic	Tay ben closer to Tay pel than to Ran pel
-0.078	-51.756	0	208435	243910	13196	Rannoch_pelagic	Tay_pelagic	Rannoch_benthic	naSealga_benthic	Ran pel closer to Ran ben than to Tay pel
-0.082	-53.996	0	208685	246054	13249	Rannoch_pelagic	Tay_pelagic	Rannoch_piscivorous	naSealga_benthic	Ran pel closer to Ran pisc than to Tay pel

0.076	51.791	0	239425	205505	13135	Rannoch_pelagic	Tay_pelagic	Tay_benthic	naSealga_benthic	Tay pel closer to Tay ben than to Ran pel
-0.189	-123.630	0	190297	278761	13156	Rannoch_piscivorous	Tay_benthic	Rannoch_benthic	naSealga_benthic	Ran pisc closer to Ran ben than to Tay ben
-0.076	-50.335	0	206970	240977	13244	Rannoch_piscivorous	Tay_benthic	Rannoch_pelagic	naSealga_benthic	Ran pisc closer to Ran pel than to Tay ben
0.085	58.465	0	238438	201115	13138	Rannoch_piscivorous	Tay_benthic	Tay_pelagic	naSealga_benthic	Tay ben closer to Tay pel than to Ran pisc
-0.210	-132.601	0	188369	288596	13201	Rannoch_piscivorous	Tay_pelagic	Rannoch_benthic	naSealga_benthic	Ran pisc closer to Ran ben than to Tay pel
-0.081	-51.426	0	209160	246054	13282	Rannoch_piscivorous	Tay_pelagic	Rannoch_pelagic	naSealga_benthic	Ran pisc closer to Ran pel than to Tay pel
0.057	35.853	0	238438	212652	13126	Rannoch_piscivorous	Tay_pelagic	Tay_benthic	naSealga_benthic	Tay pel closer to Tay ben than to Ran pisc
-0.028	-19.974	0	200277	211868	12967	Tay_benthic	Tay_pelagic	Rannoch_benthic	naSealga_benthic	Tay ben closer to Ran ben than to Tay pel
-0.007	-5.205	0	202620	205505	13090	Tay_benthic	Tay_pelagic	Rannoch_pelagic	naSealga_benthic	Tay ben closer to Ran pel than to Tay pel
-0.028	-19.578	0	201115	212652	13023	Tay_benthic	Tay_pelagic	Rannoch_piscivorous	naSealga_benthic	Tay ben closer to Ran pisc than to Tay pel

Table S4.5 – Metric principal component loadings and explained variance

metric	variance explained by PC1 [%]	Awe PC1 loading	Dùghaill PC1 loading	na Sealga PC1 loading	Rannoch PC1 loading	Tay PC1 loading	variance explained by PC2 [%]	variance explained by PC3 [%]	variance explained by PC4 [%]	variance explained by PC5 [%]	Comment
FST	31.26	0.42	0.51	0.26	0.51	0.48	19.53	17.47	16.09	15.65	
AFD	34.51	0.43	0.50	0.29	0.50	0.49	19.01	16.72	14.94	14.82	
Tajima's D (benthic)	67.88	-0.39	-0.39	-0.47	-0.48	-0.49	14.22	9.53	4.86	3.50	
Tajima's D (pelagic)	68.33	-0.39	-0.39	-0.48	-0.48	-0.48	14.40	9.10	4.50	3.66	
Tajima's D (benthic + pelagic)	66.06	-0.28	-0.26	-0.34	-0.34	-0.34	11.74	5.65	4.83	3.25	benthic loadings pelagic loadings
		-0.28	-0.26	-0.35	-0.34	-0.34					
dXY	64.53	-0.43	-0.45	-0.45	-0.46	-0.45	9.92	9.34	8.46	7.74	
pi (benthic)	67.63	-0.39	-0.39	-0.47	-0.48	-0.49	14.38	9.53	4.90	3.56	
pi (pelagic)	68.24	-0.39	-0.39	-0.48	-0.48	-0.48	14.42	9.13	4.53	3.69	

Table S4.6 – Correlation tests between metric pairs. *PCAngsd differentiation scan values are untransformed PC3 scores.

PC1 of metric 1 (z-transformed)	PC1 of metric 2 (z-transformed)*	correlation coefficient	P value	t-statistic	degrees of freedom (df)
FST	AFD	0.87	< 2.2E-16	676.9	150990
FST	Tajima's D (benthic)	0.21	< 2.2E-16	82.4	150990
FST	Tajima's D (pelagic)	0.20	< 2.2E-16	78.7	150990
FST	Tajima's D (benthic + pelagic)	0.21	< 2.2E-16	82.1	150990
FST	PCAngsd differentiation scan	0.25	< 2.2E-16	100.8	150990
AFD	Tajima's D (benthic)	-0.08	< 2.2E-16	-29.7	150990
AFD	Tajima's D (pelagic)	-0.08	< 2.2E-16	-30.2	150990
AFD	Tajima's D (benthic + pelagic)	-0.08	< 2.2E-16	-30.1	150990
AFD	PCAngsd differentiation scan	0.24	< 2.2E-16	96.3	150990
Tajima's D (benthic)	PCAngsd differentiation scan	0.14	< 2.2E-16	53.2	150990
Tajima's D (benthic)	Tajima's D (pelagic)	0.94	< 2.2E-16	1061.9	150990
Tajima's D (benthic)	Tajima's D (benthic + pelagic)	0.98	< 2.2E-16	2186.0	150990
Tajima's D (pelagic)	PCAngsd differentiation scan	0.14	< 2.2E-16	54.7	150990
Tajima's D (pelagic)	Tajima's D (benthic + pelagic)	0.98	< 2.2E-16	2194.9	150990
Tajima's D (benthic + pelagic)	PCAngsd differentiation scan	0.14	< 2.2E-16	54.9	150990
dXY	FST	0.14	< 2.2E-16	56.2	150990
dXY	AFD	-0.20	< 2.2E-16	-79.4	150990
dXY	pi (benthic)	0.88	< 2.2E-16	729.5	150990
dXY	pi (pelagic)	0.89	< 2.2E-16	754.3	150990
dXY	Tajima's D (benthic)	0.88	< 2.2E-16	733.7	150990
dXY	Tajima's D (pelagic)	0.89	< 2.2E-16	760.3	150990
dXY	Tajima's D (benthic + pelagic)	0.90	< 2.2E-16	800.1	150990
dXY	PCAngsd differentiation scan	0.10	< 2.2E-16	39.0	150990

pi (benthic)	FST	0.20	< 2.2E-16	80.8	150990
pi (benthic)	AFD	-0.08	< 2.2E-16	-31.1	150990
pi (benthic)	Tajima's D (benthic)	1.00	< 2.2E-16	8303.7	150990
pi (benthic)	Tajima's D (pelagic)	0.94	< 2.2E-16	1053.5	150990
pi (benthic)	Tajima's D (benthic + pelagic)	0.98	< 2.2E-16	2112.4	150990
pi (benthic)	PCAngsd differentiation scan	0.13	< 2.2E-16	52.5	150990
pi (benthic)	pi (pelagic)	0.94	< 2.2E-16	1058.2	150990
pi (pelagic)	FST	0.19	< 2.2E-16	76.6	150990
pi (pelagic)	AFD	-0.08	< 2.2E-16	-31.9	150990
pi (pelagic)	Tajima's D (benthic)	0.94	< 2.2E-16	1046.1	150990
pi (pelagic)	Tajima's D (pelagic)	1.00	< 2.2E-16	7564.6	150990
pi (pelagic)	Tajima's D (benthic + pelagic)	0.98	< 2.2E-16	2086.2	150990
pi (pelagic)	PCAngsd differentiation scan	0.14	< 2.2E-16	53.9	150990

Table S4.7 – Overlap of outlier genes with QTLs. Markers are from linkage maps. CI = confidence interval of QTL peak

Marker	ChromosomeID	Left CI	Right CI	Trait	Class	Species	outlier gene start	outlier gene end
grid1 (chr11)							5363821	5819813
Sal_BL_003_peak_f	NC_027327.1	2764507	12509665	Body length	Morphology	<i>Salvelinus alpinus</i>		
Sal_GR_004_right_f	NC_027327.1	0	7880658	Growth rate	Life history	<i>Salvelinus alpinus</i>		
Ssa_SM_004_peak_f	NC_027327.1	1779841	10848671	Sexual maturation	Life history	<i>Salmo salar</i>		
Omy_ED_006_msat_peak_f	NC_027327.1	0	15895419	Embryonic developmental rate	Life history	<i>Oncorhynchus mykiss</i>		
Omy_SM_029_msat_peak_f	NC_027327.1	0	15895535	Sexual maturation	Life history	<i>Oncorhynchus mykiss</i>		
grid1 (chr21)							46723466	47146887
Sal_HS_044_peak_f	NC_027318.1	34504567	53994883	Head shape	Morphology	<i>Salvelinus alpinus</i>		
Omy_BW_004_msat_peak_f	NC_027318.1	36898507	52928029	Body weight	Morphology	<i>Oncorhynchus mykiss</i>		
Sal_CF_054_peak_f	NC_027318.1	46651346	66141662	Condition factor	Morphology	<i>Salvelinus alpinus</i>		
diaph2							49681477	50365072
Sal_BW_043_EST	NC_027308.1	47482927	66973243	Body weight	Morphology	<i>Salvelinus alpinus</i>		
mdga2							47198466	47576468
Ots_BL_008_RAD_f	NC_027314.1	4666460	66666460	Body length	Morphology	<i>Oncorhynchus tshawytscha</i>		
ptprn2							26036962	26309592
Omy_BS_016_RAD_f	NC_027318.1	22396232	82978132	Body shape	Morphology	<i>Oncorhynchus mykiss</i>		

Table S5.1 – Introgressed genes under selection in benthic charr. chr = chromosome; midPos = central position of window; mu = selection statistic μ ; ref_class = class of gene; gene_symbol = zebrafish gene symbols; numABBA = number of observed ABBA patterns; numBABA = number of observed BABA patterns; numBinomialP = binomial P value; D = introgression statistic D

chr	midPos	mu	ref_class	gene_symbol	numABBA	numBABA	numBinomialP	D
NC_036842.1	36657301	12.38	pseudogene	NA	348	162	5.90913E-17	0.76
NC_036842.1	36665575	14.89	pseudogene	NA	348	162	5.90913E-17	0.76
NC_036842.1	68881679	12.56	gene	ankrd34c	428	204	1.54189E-19	0.75
NC_036842.1	86323882	15.19	gene	cftr	219	61	2.26032E-22	0.83
NC_036844.1	12524979	12.48	gene	tnc	350	138	1.3712E-22	0.83
NC_036844.1	23999851	12.9	gene	igsf9	348	117	5.55747E-28	0.72
NC_036844.1	34264300	13.26	gene	arl3	298	83	7.4916E-30	0.84
NC_036844.1	34270996	12.77	gene	rgs14	298	83	7.4916E-30	0.84
NC_036848.1	3324994	21.52	gene	ddit4	627	162	1.34246E-65	0.84
NC_036848.1	24849017	14.86	gene	diaph2	358	200	1.11228E-11	0.80
NC_036848.1	30893122	14.41	gene	NA	337	161	1.13986E-15	0.72
NC_036848.1	30998085	12.98	gene	NA	337	161	1.13986E-15	0.72
NC_036849.1	28482401	14.35	gene	colec11	159	101	0.000193903	0.72
NC_036849.1	28500610	12.65	gene	myo6	159	101	0.000193903	0.72
NC_036851.1	42321170	14.31	gene	drd1b	392	153	1.74299E-25	0.80
NC_036851.1	42327668	19.95	gene	otop1	392	153	1.74299E-25	0.80
NC_036851.1	42335290	14.02	gene	tmem128	392	153	1.74299E-25	0.80
NC_036851.1	42450583	12.56	gene	zbtb49	392	153	1.74299E-25	0.80
NC_036853.1	27474920	15.13	gene	ptprf	276	121	2.44056E-15	0.72
NC_036859.1	30025105	12.23	gene	ppm1l	577	210	1.14604E-40	0.82
NC_036865.1	2735929	12.24	gene	zdhhc6	371	185	1.18786E-15	0.75
NC_036866.1	19155491	12.21	gene	cdh13	353	148	1.15894E-20	0.72

NC_036868.1	9838387	19.44	gene	lats1	1	2	0.5	0.81
NC_036868.1	9965260	72.24	gene	nup43	1	2	0.5	0.81
NC_036868.1	9968785	85.75	gene	pcmt	1	2	0.5	0.81
NC_036868.1	10017065	67.05	gene	map2k5	1	2	0.5	0.81
NC_036868.1	10028012	48.56	gene	moxd1	1	2	0.5	0.81
NC_036868.1	10084755	19.72	pseudogene	NA	1	2	0.5	0.81
NC_036868.1	10086224	13.98	gene	taar1	1	2	0.5	0.81

Table S5.2 – Introgressed genes under selection in pelagic charr. chr = chromosome; midPos = central position of window; mu = selection statistic μ ; ref_class = class of gene; gene_symbol = zebrafish gene symbols; numABBA = number of observed ABBA patterns; numBABA = number of observed BABA patterns; numBinomialP = binomial P value; D = introgression statistic D

chr	midPos	mu	ref_class	gene_symbol	numABBA	numBABA	numBinomialP	D
NC_036842.1	57851311	12.1	gene	srpk3	308	78	9.93304E-34	0.82
NC_036842.1	57922668	12.04	gene	pus7l	308	78	9.93304E-34	0.82
NC_036842.1	68318475	15.3	gene	scaper	436	192	4.46198E-23	0.74
NC_036842.1	68863225	12.37	gene	ankrd34c	428	204	1.54189E-19	0.75
NC_036844.1	11820879	12.27	gene	s100a1	339	145	2.93094E-19	0.74
NC_036848.1	21161672	12.34	gene	rapgef2	545	164	6.49096E-49	0.74
NC_036856.1	7212598	14.32	gene	fam163a	385	180	1.9973E-18	0.72
NC_036859.1	30018500	12.9	gene	ppm1l	577	210	1.14604E-40	0.82
NC_036868.1	9906907	53.25	gene	katna1	1	2	0.5	0.81
NC_036868.1	9907083	51.6	gene	lats1	1	2	0.5	0.81
NC_036868.1	9965371	81.76	gene	nup43	1	2	0.5	0.81
NC_036868.1	9972343	50.55	gene	pcmt	1	2	0.5	0.81
NC_036868.1	9980843	61.73	gene	lrp11	1	2	0.5	0.81
NC_036868.1	10046598	23.1	gene	moxd1	1	2	0.5	0.81
NC_036868.1	10084730	13.89	pseudogene	NA	1	2	0.5	0.81
NC_036868.1	10086207	13.51	gene	taar1	1	2	0.5	0.81
NC_036872.1	36061645	12.3	gene	scarf2	109	42	2.38237E-08	0.73
NC_036872.1	36089941	12.84	pseudogene	NA	109	42	2.38237E-08	0.73

Table S5.3 – Adaptive introgression candidate genes. The shown genes were under selection in both benthic and pelagic charr and were subject to introgression between benthic and pelagic charr. chr = chromosome; midPos = central position of window; mu = selection statistic μ ; ref_class = class of gene; gene_symbol = zebrafish gene symbols; numABBA = number of observed ABBA patterns; numBABA = number of observed BABA patterns; numBinomialP = binomial P value; D = introgression statistic D

chr	midPos	mu	ref_class	gene_symbol	numABBA	numBABA	numBinomialP	D
NC_036842.1	36657301	12.38	pseudogene	NA	348	162	5.90913E-17	0.76
NC_036842.1	36665575	14.89	pseudogene	NA	348	162	5.90913E-17	0.76
NC_036842.1	68881679	12.56	gene	ankrd34c	428	204	1.54189E-19	0.75
NC_036848.1	30893122	14.41	gene	NA	337	161	1.13986E-15	0.72
NC_036848.1	30998085	12.98	gene	NA	337	161	1.13986E-15	0.72
NC_036859.1	30025105	12.23	gene	ppm1l	577	210	1.14604E-40	0.82
NC_036868.1	9838387	19.44	gene	lats1	1	2	0.5	0.81
NC_036868.1	9965260	72.24	gene	nup43	1	2	0.5	0.81
NC_036868.1	9968785	85.75	gene	pcmt	1	2	0.5	0.81
NC_036868.1	10028012	48.56	gene	moxd1	1	2	0.5	0.81
NC_036868.1	10084755	19.72	pseudogene	NA	1	2	0.5	0.81
NC_036868.1	10086224	13.98	gene	taar1	1	2	0.5	0.81

Appendix C: Additional manuscript under review



1

2 Main Manuscript for

3 Genomic basis of fishing-associated selection varies with population
4 density.

5

6 Amélie Crespel^{1,2*}, Kevin Schneider¹, Toby Miller¹, Anita Rácz^{1,3}, Arne Jacobs^{1,4}, Jan Lindström¹,
7 Kathryn R. Elmer¹, Shaun S. Killen¹

8 ¹University of Glasgow, Institute of Biodiversity, Animal Health and Comparative Medicine,
9 Glasgow, G12 8QQ, UK.

10 ²Department of Biology, University of Turku, 20500 Turku, Finland

11 ³Department of Genetics, Eötvös Loránd University, Budapest, H-1117, Hungary.

12 ⁴Department of Natural Resources, Cornell University, 14853, Ithaca, NY, USA.

13 * Amélie Crespel (Department of Biology, University of Turku, 20500 Turku, Finland,
14 amelie.crespel@gmail.com), Shaun Killen (Institute of Biodiversity, Animal Health and Comparative
15 Medicine, University of Glasgow, Glasgow, G12 8QQ, UK, shaun.killen@glasgow.ac.uk)

16 Email: amelie.crespel@gmail.com, shaun.killen@glasgow.ac.uk

17 ORCID ID: Amélie Crespel 0000-0002-6351-9008

18

19

20 Classification

21 Biological Sciences, Genetics

22 Keywords

23 Fisheries, density-dependent effects, sequencing, anthropogenic effects, environmental change

24

25

26 **Author Contributions**

27 Conceptualization: AC, JL, SSK; Methodology: AC, AJ, JL, KRE, SSK; Software: KS, AJ; Formal
 28 analysis: AC, KS, TM; Resources: AC, AR, SSK; Writing – original draft: AC, KS; Writing – review
 29 and editing: AC, KS, AJ, JL, KRE, SSK; Supervision: AC, JL, KRE, SSK; Project administration:
 30 AC, SSK; Funding acquisition: AC, SSK. All authors approved the manuscript before submission.

31 **This PDF file includes:**

32 Main Text
 33 Figures 1 to 5
 34 Tables 1 to 1
 35 Suppl. Figures S1 to S2
 36 Suppl. Tables S1 to S3
 37

38 **Abstract (250)**

39 Fisheries induce strong selective pressure on natural populations but the genetic effects of fishing
 40 remain unclear. Crucially, we lack knowledge of how capture-associated selection, and its
 41 interaction with reductions in population density caused by fishing, can potentially shift which
 42 genes are under selection. Using experimental fish reared at two densities and repeatedly
 43 harvested by trawling, we show consistent phenotypic selection on growth, metabolism, and
 44 social behaviour regardless of density. However, the specific genes under selection – mainly
 45 related to brain function and neurogenesis – depended on density. This interaction between direct
 46 fishing selection and density fundamentally alters genomic responses to harvest. The
 47 evolutionary consequences of fishing are likely context-dependent, varying as exploited
 48 populations decline, highlighting the need to consider environmental factors when predicting
 49 effects of human-induced evolution.

50 **Significance Statement (50-120)**

51 Fisheries-associated selection is recognized as one of the strongest drivers of contemporary
 52 evolution in natural populations. The results of this study show that while commercial fishing
 53 techniques consistently remove fish with traits associated with growth, metabolism, and
 54 behaviour, the exact genes under selection differ depending on the density of the targeted fish
 55 population. This finding suggests that different fish populations of varying sizes will respond
 56 differently to fishing selection at the genetic level. Furthermore, as a population is fished over

57 time, the genes under selection will change as the population diminishes. This could have
58 profound impacts on population resilience. This study highlights the relative importance of
59 selection and ecological effects on harvested fish populations and identify the genes under
60 selection.

Appendix D: Additional published manuscripts

Molecular Phylogenetics and Evolution 153 (2020) 106950



Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev

Cryptic species in White Cloud Mountain minnow, *Tanichthys albonubes*: Taxonomic and conservation implications



Chao Li^{a,b}, Shuying Jiang^a, Kevin Schneider^b, Jinjin Jin^a, Hungdu Lin^c, Junjie Wang^a, Kathryn R. Elmer^b, Jun Zhao^{a,*}

^a Guangdong Provincial Key Laboratory for Healthy and Safe Aquaculture, Guangdong Provincial Engineering Technology Research Center for Environmentally Friendly Aquaculture, Guangzhou Key Laboratory of Subtropical Biodiversity and Biomonitoring, School of Life Science, South China Normal University, Guangzhou 510631, China

^b Institute of Biodiversity, Animal Health and Comparative Medicine, College of Medical, Veterinary and Life Sciences, University of Glasgow, Glasgow G12 8QQ, Scotland, UK

^c The Affiliated School of National Tainan First Senior High School, Tainan, Taiwan

ARTICLE INFO

Keywords:
Tanichthys albonubes
Tanichthys micagemmae
 Cryptic species
 Species delimitation
 Conservation

ABSTRACT

Cryptic species describe two or more species that had mistakenly been considered to be a single species, a phenomenon that has been found throughout the tree of life. Recognizing cryptic species is key to estimating the real biodiversity of the world and understanding evolutionary processes. Molecular methods present an unprecedented opportunity for biologists to question whether morphologically similar populations are actually cryptic species. The minnow *Tanichthys albonubes* is a critically endangered freshwater fish and was classified as a second-class state-protected animal in China. Previous studies have revealed highly divergent lineages with similar morphological characters in this species. Herein, we tested for cryptic species across the ranges of all known wild populations of this minnow. Using multilocus molecular (one mitochondrial gene, two nuclear genes and 13 microsatellite loci) and morphological data for 230 individuals from eight populations, we found deep genetic divergence among these populations with subtle morphological disparity. Morphological examination found variance among these populations in the number of branched anal-fin rays. Based on genetic data, we inferred eight monophyletic groups that were well supported by haplotype network and population clustering analyses. Species delimitation methods suggested eight putative species in the *T. albonubes* complex. Molecular dating suggested that these cryptic species diverged in the period from the Pliocene to the Pleistocene. Based on these findings, we propose the existence of seven cryptic species in the *T. albonubes* complex. Our results highlight the need for a taxonomic revision of *Tanichthys*. What is more, the conservation status of and conservation strategies for the *T. albonubes* complex should be reassessed as soon as possible.



REGULAR PAPER

Native bighead carp *Hypophthalmichthys nobilis* and silver carp *Hypophthalmichthys molitrix* populations in the Pearl River are threatened by Yangtze River introductions as revealed by mitochondrial DNA

Chao Li^{1,2*} | Junjie Wang^{1*} | Jiaqi Chen¹ | Kevin Schneider² |
Radhakrishnan K. Veetil³ | Kathryn R. Elmer² | Jun Zhao¹

¹Guangzhou Key Laboratory of Subtropical Biodiversity and Biomonitoring, Guangdong Provincial Key Laboratory for Healthy and Safe Aquaculture, Guangdong Provincial Engineering Technology Research Center for Environmentally Friendly Aquaculture, School of Life Science, South China Normal University, Guangzhou, China

²Institute of Biodiversity, Animal Health and Comparative Medicine, College of Medical, Veterinary and Life Sciences, University of Glasgow, Glasgow, Scotland

³College of Fisheries, Central Agricultural University, Agartala, India

Correspondence

Jun Zhao, Guangzhou Key Laboratory of Subtropical Biodiversity and Biomonitoring, Guangdong Provincial Key Laboratory for Healthy and Safe Aquaculture, Guangdong Provincial Engineering Technology Research Center for Environmentally Friendly Aquaculture, School of Life Science, South China Normal University, Guangzhou 510631, China.
Email: zhaojun@scnu.edu.cn

Funding information

China-ASEAN Maritime Cooperation Fund, Grant/Award Number: CAMC-2018F; Guangdong Provincial Government, Grant/Award Number: 2018010162; Key Project of National Science, Technology Basic Conditions Platform Work, Grant/Award Number: 2005DKA21402; South China Normal University; Special Fund for Agro-scientific Research in the Public Interest, Grant/Award Number: 201303048; China Scholarship Council, Grant/Award Number: 201906750010; South China Normal University

Abstract

Bighead carp *Hypophthalmichthys nobilis* and silver carp *Hypophthalmichthys molitrix* have been two economically important aquaculture species in China for centuries. In the past decades, bighead and silver carp have been introduced from the Yangtze River to many river systems in China, including the Pearl River, in annual, large-scale, stocking activities to enhance wild fisheries. Nonetheless, few studies have assessed the ecological or genetic impacts of such introductions on native conspecific fish populations. An mtDNA D-loop segment of 978 bp from 213 bighead carp samples from 9 populations and a 975 bp segment from 204 silver carp samples from 10 populations were obtained to evaluate genetic diversity and population integrity. Results from a haplotype network analysis revealed that most haplotypes of the Pearl River clustered with those of Yangtze River origin and only a small proportion were distinct, suggesting that both the native Pearl River bighead and silver carp populations are currently dominated by genetic material from the Yangtze River. The genetic diversity of Pearl River populations is high in both species because of this inter-population gene flow, but the diversity of native Pearl River populations is low. To preserve the native genetic diversity, stocking of non-native fingerlings should cease immediately and native Pearl River bighead and silver carp fish farms should be established. This research demonstrates the danger to native biodiversity across China because of the substantial, ongoing stock-enhancement activities without prior genetic assessment.

KEYWORDS

bighead carp, domestic introduction, genetic diversity, the Pearl River, silver carp, stock enhancement

*These authors contributed equally to this work.

References

- Aardema, M. L., Stiassny, M. L., & Alter, S. E. (2020). Genomic analysis of the only blind cichlid reveals extensive inactivation in eye and pigment formation genes. *Genome Biology and Evolution*, *evaa144*, DOI: <https://doi.org/10.1093/gbe/evaa144>.
- Adams, C. E., & Huntingford, F. A. (2004). Incipient speciation driven by phenotypic plasticity? Evidence from sympatric populations of Arctic charr. *Biological Journal of the Linnean Society*, *81*(4), 611-618.
- Adams, C. E., Fraser, D., Huntingford, F. A., Greer, R. B., Askew, C. M., & Walker, A. F. (1998). Trophic polymorphism amongst arctic charr from Loch Rannoch, Scotland. *Journal of Fish Biology*, *52*(6), 1259-1271.
- Adams, C. E., Fraser, D., Wilson, A. J., Alexander, G., Ferguson, M. M., & Skúlason, S. (2007). Patterns of phenotypic and genetic variability show hidden diversity in Scottish Arctic charr. *Ecology of Freshwater Fish*, *16*(1), 78-86.
- Afanasyeva, A., Bockwoldt, M., Cooney, C. R., Heiland, I., & Gossmann, T. I. (2018). Human long intrinsically disordered protein regions are frequent targets of positive selection. *Genome Research*, *28*, 975-982.
- Ahi, E. P. (2016). Studies of craniofacial gene expression during embryonic development in divergent Arctic charr (*Salvelinus alpinus*) morphs. (Doctoral dissertation). University of Iceland, Reykjavík, Iceland.
- Ahi, E. P., Kapralova, K. H., Pálsson, A., Maier, V. H., Gudbrandsson, J., Snorrason, S. S., ... & Franzdóttir, S. R. (2014). Transcriptional dynamics of a conserved gene expression network associated with craniofacial divergence in Arctic charr. *EvoDevo*, *5*, 40.
- Alachiotis, N., & Pavlidis, P. (2018). RAIiSD detects positive selection based on multiple signatures of a selective sweep and SNP vectors. *Communications Biology*, *1*(1), 1-11.

- Alachiotis, N., Stamatakis, A., & Pavlidis, P. (2012). OmegaPlus: a scalable tool for rapid detection of selective sweeps in whole-genome datasets. *Bioinformatics*, 28(17), 2274-2275.
- Alcántata-Rodríguez, J. A., Ciroso-Pérez, J., Ortega-Mayagoitia, E., Serrania-Soto, C. R., & Piedra-Ibarra, E. (2012). Local adaptation in populations of a *Brachionus* group *plicatilis* cryptic species inhabiting three deep crater lakes in Central Mexico. *Freshwater Biology*, 57(4), 728-740.
- Alexander, G. D., & Adams, C. E. (2000). The phenotypic diversity of Arctic charr, *Salvelinus alpinus* (Salmonidae) in Scotland and Ireland. *Aqua*, 4, 77-88.
- Alexander, G. D., & Adams, C. E. (2004). Exposure to a common environment erodes inherited between-population trophic morphology differences in Arctic charr. *Journal of Fish Biology*, 64(1), 253-257.
- Altenhoff, A. M., Schneider, A., Gonnet, G. H., & Dessimoz, C. (2010). OMA 2011: orthology inference among 1000 complete genomes. *Nucleic Acids Research*, 39(suppl_1), D289-D294.
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215(3), 403-410.
- Anaya-Rojas, J. M., Brunner, F. S., Sommer, N., Seehausen, O., Eizaguirre, C., & Matthews, B. (2016). The association of feeding behaviour with the resistance and tolerance to parasites in recently diverged sticklebacks. *Journal of Evolutionary Biology*, 29(11), 2157-2167.
- Andrews, S. (2010). FastQC: A quality control tool for high throughput sequence data. Available online at: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>
- Arbour, J. H., & López-Fernández, H. (2016). Continental cichlid radiations: functional diversity reveals the role of changing ecological opportunity in the Neotropics. *Proceedings of the Royal Society B*, 283, 20160556.

- Bailey, S. F., Rodrigue, N., & Kassen, R. (2015). The effect of selection environment on the probability of parallel evolution. *Molecular Biology and Evolution*, 32:1436-1448.
- Baillie, S. M., Muir, A. M., Hansen, M. J., Krueger, C. C., & Bentzen, P. (2016). Genetic and phenotypic variation along an ecological gradient in lake trout *Salvelinus namaycush*. *BMC Evolutionary Biology*, 16(1), 1-16.
- Bamshad, M., & Wooding, S. P. (2003). Signatures of natural selection in the human genome. *Nature Reviews Genetics*, 4(2), 99-110.
- Barrett, R. D., & Hoekstra, H. E. (2011). Molecular spandrels: tests of adaptation at the genetic level. *Nature Reviews Genetics*, 12(11), 767-780.
- Barrett, R. D., Laurent, S., Mallarino, R., Pfeifer, S. P., Xu, C. C., Foll, M., ... & Hoekstra, H. E. (2019). Linking a mutation to survival in wild mice. *Science*, 363(6426), 499-504.
- Bay, R. A., Taylor, E. B., & Schluter, D. (2019). Parallel introgression and selection on introduced alleles in a native species. *Molecular Ecology*, 28(11), 2802-2813.
- Beaumont, M. A. (2005). Adaptation and speciation: what can F_{st} tell us? *Trends in Ecology & Evolution*, 20(8), 435-440.
- Beck, S. V., Räsänen, K., Ahi, E. P., Kristjánsson, B. K., Skúlason, S., Jónsson, Z. O., & Leblanc, C. A. (2019). Gene expression in the phenotypically plastic Arctic charr (*Salvelinus alpinus*): a focus on growth and ossification at early stages of development. *Evolution & Development*, 21, 16.
- Bence, K. K. (Ed.). (2013). Protein tyrosine phosphatase control of metabolism. New York, NY: Springer.
- Benderoth, M., Textor, S., Windsor, A. J., Mitchell-Olds, T., Gershenzon, J., & Kroymann, J. (2006). Positive selection driving diversification in plant secondary metabolism. *Proceedings of the National Academy of Sciences*, 103, 9118-9123.

- 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(1), 289-300.
- Bernatchez, L., & Wilson, C. C. (1998). Comparative phylogeography of Nearctic and Palearctic fishes. *Molecular Ecology*, 7, 431-452.
- Bernatchez, L., Renaut, S., Whiteley, A. R., Derome, N., Jeukens, J., Landry, L., ... & St-Cyr, J. (2010). On the origin of species: insights from the ecological genomics of lake whitefish. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365:1783-1800.
- Bernatchez, S., Laporte, M., Perrier, C., Sirois, P., & Bernatchez, L. (2016). Investigating genomic and phenotypic parallelism between piscivorous and planktivorous lake trout (*Salvelinus namaycush*) ecotypes by means of RAD seq and morphometrics analyses. *Molecular Ecology*, 25:4773-4792.
- Berner, D., Grandchamp, A. C., & Hendry, A. P. (2009). Variable progress toward ecological speciation in parapatry: stickleback across eight lake-stream transitions. *Evolution*, 63:1740-1753.
- Berner, D., Salzburger, W. (2015). The genomics of organismal diversification illuminated by adaptive radiations. *Trends in Genetics*, 31:491-499.
- Berner, D. (2019). Allele frequency difference AFD—an intuitive alternative to F_{ST} for quantifying genetic population differentiation. *Genes*, 10(4), 308.
- Bertolotti, A. C., Layer, R. M., Gundappa, M. K., Gallagher, M. D., Pehlivanoglu, E., Nome, T., ... & Mulugeta, T. (2020). The structural variation landscape in 492 Atlantic salmon genomes. *bioRxiv*, DOI: <https://doi.org/10.1101/2020.05.16.099614>.
- Bian, X., Wasserfall, C., Wallstrom, G., Wang, J., Wang, H., Barker, K., ... & LaBaer, J. (2017). Tracking the antibody immunome in type 1 diabetes using protein arrays. *Journal of Proteome Research*, 16(1), 195-203.

- Bielawski, J. P., & Yang, Z. (2001). Positive and negative selection in the DAZ gene family. *Molecular Biology and Evolution*, 18, 523-529.
- Bierne, N., Roze, D., & Welch, J. J. (2013). Pervasive selection or is it...? Why are FST outliers sometimes so frequent?. *Molecular Ecology*, 22(8), 2061-2064.
- Bione, S., Sala, C., Manzini, C., Arrigo, G., Zuffardi, O., Banfi, S., ... & Ballabio, A. (1998). A human homologue of the *Drosophila melanogaster* diaphanous gene is disrupted in a patient with premature ovarian failure: evidence for conserved function in oogenesis and implications for human sterility. *The American Journal of Human Genetics*, 62(3), 533-541.
- Birchler, J. A., & Veitia, R. A. (2012). Gene balance hypothesis: connecting issues of dosage sensitivity across biological disciplines. *Proceedings of the National Academy of Sciences*, 109, 14746-14753.
- Bisschop, K., Mortier, F., Etienne, R. S., & Bonte, D. (2019). Transient local adaptation and source–sink dynamics in experimental populations experiencing spatially heterogeneous environments. *Proceedings of the Royal Society B*, 286(1905), 20190738.
- Bleidorn, C. (2016). Third generation sequencing: technology and its potential impact on evolutionary biodiversity research. *Systematics and Biodiversity*, 14(1), 1-8.
- Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15), 2114-2120.
- Bolnick, D. I., & Lau, O. L. (2008). Predictable patterns of disruptive selection in stickleback in postglacial lakes. *The American Naturalist*, 172(1), 1-11.
- Bolnick, D. I., Barrett, R. D., Oke, K. B., Rennison, D. J., & Stuart, Y. E. (2018). (Non) parallel evolution. *Annual Review of Ecology, Evolution, and Systematics*, 49, 303-330.

- Bonferroni, C. (1936). Teoria statistica delle classi e calcolo delle probabilit . *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, 8, 3-62.
- Bonin, A. (2008). Population genomics: a new generation of genome scans to bridge the gap with functional genomics. *Molecular Ecology*, 17(16), 3583-3584.
- Borchel, A., Verleih, M., K hn, C., Rebl, A., & Goldammer, T. (2019). Evolutionary expression differences of creatine synthesis-related genes: Implications for skeletal muscle metabolism in fish. *Scientific Reports*, 9, 5429.
- Borchel, A., Verleih, M., Rebl, A., K hn, C., & Goldammer, T. (2014). Creatine metabolism differs between mammals and rainbow trout (*Oncorhynchus mykiss*). *SpringerPlus*, 3, 510.
- Borowsky, R., & Cohen, D. (2013). Genomic consequences of ecological speciation in *Astyanax* cavefish. *PloS One*, 8(11), e79903.
- Braissant, O., Bachmann, C., & Henry, H. (2007). Expression and function of AGAT, GAMT and CT1 in the mammalian brain. In *Creatine and creatine kinase in health and disease* (pp. 67-81). Springer, Dordrecht.
- Brawand, D., Wagner, C. E., Li, Y. I., Malinsky, M., Keller, I., Fan, S., ... & Turner-Maier, J. (2014). The genomic substrate for adaptive radiation in African cichlid fish. *Nature*, 513(7518), 375-381.
- Brown, W. L., & Wilson, E. O. (1956). Character displacement. *Systematic Zoology*, 5(2), 49-64.
- Browning, S. R., & Browning, B. L. (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *The American Journal of Human Genetics*, 81(5), 1084-1097.
- Bruneaux, M., Johnston, S. E., Herczeg, G., Meril , J., Primmer, C. R., & Vasem gi, A. (2013). Molecular evolutionary and population genomic analysis of the nine-spined

stickleback using a modified restriction-site-associated DNA tag approach.

Molecular Ecology, 22, 565-582.

Brunet, F. G., Crollius, H. R., Paris, M., Aury, J. M., Gibert, P., Jaillon, O., ... & Robinson-Rechavi, M. (2006). Gene loss and evolutionary rates following whole-genome duplication in teleost fishes. *Molecular Biology and Evolution*, 23, 1808-1816.

Brunner, F. S., Anaya-Rojas, J. M., Matthews, B., & Eizaguirre, C. (2017). Experimental evidence that parasites drive eco-evolutionary feedbacks. *Proceedings of the National Academy of Sciences*, 114, 3678-3683.

Brunner, P. C., Douglas, M. R., Osinov, A., Wilson, C. C., & Bernatchez, L. (2001). Holarctic phylogeography of Arctic charr (*Salvelinus alpinus* L.) inferred from mitochondrial DNA sequences. *Evolution*, 55(3), 573-586.

Buerkle, C. A., & Gompert, Z. (2013). Population genomics based on low coverage sequencing: how low should we go?. *Molecular Ecology*, 22(11), 3028-3035.

Burri, R., Nater, A., Kawakami, T., Mugal, C. F., Olason, P. I., Smeds, L., ... & Hogner, S. (2015). Linked selection and recombination rate variation drive the evolution of the genomic landscape of differentiation across the speciation continuum of Ficedula flycatchers. *Genome Research*, 25(11), 1656-1665.

Butlin, R. K., Saura, M., Charrier, G., Jackson, B., André, C., Caballero, A., ... & Kemppainen, P. (2014). Parallel evolution of local adaptation and reproductive isolation in the face of gene flow. *Evolution*, 68(4), 935-949.

Calderoni, L., Rota-Stabelli, O., Frigato, E., Panziera, A., Kirchner, S., Foulkes, N. S., ... & Fuselli, S. (2016). Relaxed selective constraints drove functional modifications in peripheral photoreception of the cavefish *P. andruzzii* and provide insight into the time of cave colonization. *Heredity*, 117, 383.

Camon, E., Magrane, M., Barrell, D., Lee, V., Dimmer, E., Maslen, J., ... & Apweiler, R. (2004). The gene ontology annotation (goa) database: sharing knowledge in uniprot with gene ontology. *Nucleic Acids Research*, 32, D262-D266.

- Campagna, L., Repenning, M., Silveira, L. F., Fontana, C. S., Tubaro, P. L., & Lovette, I. J. (2017). Repeated divergent selection on pigmentation genes in a rapid finch radiation. *Science Advances*, 3(5), e1602404.
- Campana, M. G., Corvelo, A., Shelton, J., Callicrate, T. E., Bunting, K. L., Riley-Gillis, B., ... & Fleischer, R. C. (2020). Adaptive radiation genomics of two ecologically divergent Hawai 'ian honeycreepers: the 'akiapōlā 'au and the Hawai 'i 'amakihi. *Journal of Heredity*, 111(1), 21-32.
- Capella-Gutiérrez, S., Silla-Martínez, J.M., & Gabaldón, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, 25, 1972-1973.
- Carlini, D. B., & Fong, D. W. (2017). The transcriptomes of cave and surface populations of *Gammarus minus* (Crustacea: Amphipoda) provide evidence for positive selection on cave downregulated transcripts. *PLoS One*, 12(10), e0186173.
- Carmona, G. N., Nishimura, T., Schindler, C. W., Panlilio, L. V., & Notkins, A. L. (2014). The dense core vesicle protein IA-2, but not IA-2β, is required for active avoidance learning. *Neuroscience*, 269, 35-42.
- Carneiro, M., Baird, S. J., Afonso, S., Ramirez, E., Tarroso, P., Teotónio, H., ... & Ferrand, N. (2013). Steep clines within a highly permeable genome across a hybrid zone between two subspecies of the European rabbit. *Molecular Ecology*, 22(9), 2511-2525.
- Carruthers, M., Yurchenko, A., Augley, J. J., Adams, C. E., Herzyk, P., Elmer, K. R. (2018). De novo transcriptome assembly, annotation and comparison of four ecological and evolutionary model salmonid fish species. *BMC Genomics*, 19, 32.
- Case, T. J., & Taper, M. L. (2000). Interspecific competition, environmental gradients, gene flow, and the coevolution of species' borders. *The American Naturalist*, 155(5), 583-605.

- Castiglione, G. M., Schott, R. K., Hauser, F. E., & Chang, B. S. (2018). Convergent selection pressures drive the evolution of rhodopsin kinetics at high altitudes via nonparallel mechanisms. *Evolution*, 72, 170-186.
- Castoe, T. A., de Koning, A. J., Kim, H. M., Gu, W., Noonan, B. P., Naylor, G., ... & Pollock, D. D. (2009). Evidence for an ancient adaptive episode of convergent molecular evolution. *Proceedings of the National Academy of Sciences*, 106, 8986-8991.
- Chao, T., Liu, Z., Zhang, Y., Zhang, L., Huang, R., He, L., ... & Zheng, W. (2019). Precise and rapid validation of candidate gene by allele specific knockout with CRISPR/Cas9 in wild mice. *Frontiers in Genetics*, 10, 124.
- Charlesworth, B. (2017). Haldane and modern evolutionary genetics. *Journal of Genetics*, 96(5), 773-782.
- Charlesworth, D., Charlesworth, B., & Morgan, M. T. (1995). The pattern of neutral molecular variation under the background selection model. *Genetics*, 141(4), 1619-1632.
- Chavarie, L., Howland, K., Harris, L., & Tonn, W. (2015). Polymorphism in lake trout in Great Bear Lake: intra-lake morphological diversification at two spatial scales. *Biological Journal of the Linnean Society*, 114, 109-125.
- Chavarie, L., Muir, A. M., Zimmerman, M. S., Baillie, S. M., Hansen, M. J., Nate, N. A., ... & Krueger, C. C. (2016). Challenge to the model of lake charr evolution: shallow-and deep-water morphs exist within a small postglacial lake. *Biological Journal of the Linnean Society*, 120, 578-603.
- Chen, X., Wang, J., Yue, W., Lei, S., Dobjay, S., Li, Z., & Wang, C. (2019). Integrated transcriptome provides resources and insights into the adaptive evolution of colonized brown trout (*Salmo trutta fario*) in the Tibetan Plateau. *Journal of the World Aquaculture Society*, DOI: 10.1111/jwas.12621
- Chen, H., Patterson, N., & Reich, D. (2010). Population differentiation as a test for selective sweeps. *Genome Research*, 20(3), 393-402.

- Chen, Y., Ye, W., Zhang, Y., & Xu, Y. (2015). High speed BLASTN: an accelerated MegaBLAST search tool. *Nucleic Acids Research*, 43(16), 7762-7768.
- Chikina, M., Robinson, J. D., & Clark, N. L. (2016). Hundreds of genes experienced convergent shifts in selective pressure in marine mammals. *Molecular Biology and Evolution*, 33, 2182-2192.
- Christensen, K. A., Rondeau, E. B., Minkley, D. R., Leong, J. S., Nugent, C. M., Danzmann, R. G., ... & Edwards, M. (2018). The Arctic charr (*Salvelinus alpinus*) genome and transcriptome assembly. *PLoS One*, 13(9), e0204076.
- Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., ... & Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly*, 6(2), 80-92.
- Clarke, J. M., & Schluter, D. (2011). Colour plasticity and background matching in a threespine stickleback species pair. *Biological Journal of the Linnean Society*, 102(4), 902-914.
- Conejeros, P., Phan, A., Power, M., O'Connell, M., Alekseyev, S., Salinas, I., & Dixon, B. (2014). Differentiation of sympatric Arctic char morphotypes using major histocompatibility class II genes. *Transactions of the American Fisheries Society*, 143, 586-594.
- Conesa, A., Götz, S., García-Gómez, J. M., Terol, J., Talón, M., & Robles, M. (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, 21, 3674-3676.
- Conte, G. L., Arnegard, M. E., Peichel, C. L., & Schluter, D. (2012). The probability of genetic parallelism and convergence in natural populations. *Proceedings of the Royal Society B: Biological Sciences*, 279(1749), 5039-5047.
- Cooper, W. J., Parsons, K., McIntyre, A., Kern, B., McGee-Moore, A., & Albertson, R. C. (2010). Benthic-pelagic divergence of cichlid feeding architecture was prodigious

and consistent during multiple adaptive radiations within African rift-lakes. *PloS One*, 5(3), e9551.

Pfennig, D. W., Wund, M. A., Snell-Rood, E. C., Cruickshank, T., Schlichting, C. D., & Moczek, A. P. (2010). Phenotypic plasticity's impacts on diversification and speciation. *Trends in Ecology & Evolution*, 25(8), 459-467.

Crawford, J. E., & Lazzaro, B. P. (2012). Assessing the accuracy and power of population genetic inference from low-pass next-generation sequencing data. *Frontiers in Genetics*, 3, 66.

Crespi, B., & Nosil, P. (2013). Conflictual speciation: species formation via genomic conflict. *Trends in Ecology & Evolution*, 28(1), 48-57.

Crête-Lafrenière A, Weir LK, & Bernatchez L 2012. Framing the Salmonidae family phylogenetic portrait: a more complete picture from increased taxon sampling. *PLoS One*, 7:e46662.

Cruickshank, T. E., & Hahn, M. W. (2014). Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Molecular Ecology*, 23(13), 3133-3157.

Cui, L., Yu, W. P., DeAizpurua, H. J., Schmidli, R. S., & Pallen, C. J. (1996). Cloning and characterization of islet cell antigen-related protein-tyrosine phosphatase (PTP), a novel receptor-like PTP and autoantigen in insulin-dependent diabetes. *Journal of Biological Chemistry*, 271(40), 24817-24823.

Czelusniak, J., Goodman, M., Hewett-Emmett, D., Weiss, M. L., Venta, P. J., & Tashian, R. E. (1982). Phylogenetic origins and adaptive evolution of avian and mammalian haemoglobin genes. *Nature*, 298(5871), 297-300.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., ... & McVean, G. (2011). The variant call format and VCFtools. *Bioinformatics*, 27(15), 2156-2158.

Darwin, C. (1859). On the origin of species by means of natural selection, or preservation of favoured races in the struggle for life. John Murray, London.

- Davey, J. W., & Blaxter, M. L. (2010). RADSeq: next-generation population genetics. *Briefings in Functional Genomics*, 9(5-6), 416-423.
- De Villemereuil, P., & Gaggiotti, O. E. (2015). A new FST-based method to uncover local adaptation using environmental variables. *Methods in Ecology and Evolution*, 6(11), 1248-1258.
- De Visser, J. A. G., & Krug, J. (2014). Empirical fitness landscapes and the predictability of evolution. *Nature Reviews Genetics*, 15(7), 480-490.
- Dennenmoser, S., Vamosi, S. M., Nolte, A. W., & Rogers, S. M. (2017). Adaptive genomic divergence under high gene flow between freshwater and brackish-water ecotypes of prickly sculpin (*Cottus asper*) revealed by Pool-Seq. *Molecular Ecology*, 26, 25-42.
- Derome, N., Duchesne, P., & Bernatchez, L. (2006). Parallelism in gene transcription among sympatric lake whitefish (*Coregonus clupeaformis* Mitchill) ecotypes. *Molecular Ecology*, 15(5), 1239-1249.
- Dobler, S., Dalla, S., Wagschal, V., & Agrawal, A. A. (2012). Community-wide convergent evolution in insect adaptation to toxic cardenolides by substitutions in the Na, K-ATPase. *Proceedings of the National Academy of Sciences*, 109, 13040-13045.
- Dodson, J. J., Aubin-Horth, N., Thériault, V., & Páez, D. J. (2013). The evolutionary ecology of alternative migratory tactics in salmonid fishes. *Biological Reviews*, 88, 602-625.
- Doenz, C. J., Bittner, D., Vonlanthen, P., Wagner, C. E., & Seehausen, O. (2018). Rapid buildup of sympatric species diversity in Alpine whitefish. *Ecology and Evolution*, 8, 9398-9412.
- Doenz, C. J., Krähenbühl, A. K., Walker, J., Seehausen, O., & Brodersen, J. (2019). Ecological opportunity shapes a large Arctic charr species radiation. *Proceedings of the Royal Society B*, 286(1913), 20191992.

- Douda, K., Liu, H. Z., Yu, D., Rouchet, R., Liu, F., Tang, Q. Y., ... & Reichard, M. (2017). The role of local adaptation in shaping fish-mussel coevolution. *Freshwater Biology*, 62(11), 1858-1868.
- Doudna, J. A., & Charpentier, E. (2014). The new frontier of genome engineering with CRISPR-Cas9. *Science*, 346(6213).
- Du, M., Chen, S. L., Liu, Y. H., Liu, Y., & Yang, J. F. (2011). MHC polymorphism and disease resistance to *Vibrio anguillarum* in 8 families of half-smooth tongue sole (*Cynoglossus semilaevis*). *BMC Genetics*, 12, 78.
- Dunham, M. J., Badrane, H., Ferea, T., Adams, J., Brown, P. O., Rosenzweig, F., & Botstein, D. (2002). Characteristic genome rearrangements in experimental evolution of *Saccharomyces cerevisiae*. *Proceedings of the National Academy of Sciences*, 99(25), 16144-16149.
- Dunn, O. J. (1961). Multiple comparisons among means. *Journal of the American Statistical Association*, 56:52-64.
- Durand, E. Y., Patterson, N., Reich, D., & Slatkin, M. (2011). Testing for ancient admixture between closely related populations. *Molecular Biology and Evolution*, 28(8), 2239-2252.
- Early, A. M., & Clark, A. G. (2017). Genomic signatures of local adaptation in the *Drosophila* immune response. *Fly*, 11(4), 277-283.
- Egan, S. P., Ragland, G. J., Assour, L., Powell, T. H., Hood, G. R., Emrich, S., ... & Feder, J. L. (2015). Experimental evidence of genome-wide impact of ecological selection during early stages of speciation-with-gene-flow. *Ecology Letters*, 18(8), 817-825.
- Ekblom, R., & Galindo, J. (2011). Applications of next generation sequencing in molecular ecology of non-model organisms. *Heredity*, 107(1), 1-15.
- Ellegren, H. (2014). Genome sequencing and population genomics in non-model organisms. *Trends in Ecology & Evolution*, 29(1), 51-63.

- Elmer, K. R., Fan, S., Gunter, H. M., Jones, J. C., Boekhoff, S., Kuraku, S., & Meyer, A. (2010). Rapid evolution and selection inferred from the transcriptomes of sympatric crater lake cichlid fishes. *Molecular Ecology*, 19:197-211.
- Elmer, K. R., Reggio, C., Wirth, T., Verheyen, E., Salzburger, W., Meyer, A. (2009). Pleistocene desiccation in East Africa bottlenecked but did not extirpate the adaptive radiation of Lake Victoria haplochromine cichlid fishes. *Proc Natl Acad Sci USA*, 106, 13404-13409.
- Elmer, K. R. (2016). Genomic tools for new insights to variation, adaptation, and evolution in the salmonid fishes: a perspective for charr. *Hydrobiologia*, 783(1), 191-208.
- Elmer, K. R., & Meyer, A. (2011). Adaptation in the age of ecological genomics: insights from parallelism and convergence. *Trends in Ecology & Evolution*, 26(6), 298-306.
- Elmer, K. R., Fan, S., Kusche, H., Spreitzer, M. L., Kautt, A. F., Franchini, P., & Meyer, A. (2014). Parallel evolution of Nicaraguan crater lake cichlid fishes via non-parallel routes. *Nature Communications*, 5(1), 1-8.
- Emms, D. M., & Kelly, S. (2015). OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology*, 16(1), 157.
- Emms, D. M., & Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biology*, 20(1), 1-14.
- Enard, D., Messer, P. W., & Petrov, D. A. (2014). Genome-wide signals of positive selection in human evolution. *Genome Research*, 24(6), 885-895.
- Enciso-Romero, J., Pardo-Díaz, C., Martin, S. H., Arias, C. F., Linares, M., McMillan, W. O., ... & Salazar, C. (2017). Evolution of novel mimicry rings facilitated by adaptive introgression in tropical butterflies. *Molecular Ecology*, 26(19), 5160-5172.

- Esin, E. V., Bocharova, E. S., Muge, N. S., & Markevich, G. N. (2017). Occurrence of sympatric charr groups, *Salvelinus*, Salmonidae, in the lakes of Kamchatka: a legacy of the last glaciations. *Journal of Fish Biology*, 91(2), 628-644.
- Excoffier, L., Hofer, T., & Foll, M. (2009). Detecting loci under selection in a hierarchically structured population. *Heredity*, 103(4), 285-298.
- Eyre-Walker, A. (1999). Evolutionary genomics. *Trends in Ecology & Evolution*, 14(5), 176.
- Fan, S., Elmer, K. R., & Meyer, A. (2011). Positive Darwinian selection drives the evolution of the morphology-related gene, *EPCAM*, in particularly species-rich lineages of African cichlid fishes. *Journal of Molecular Evolution*, 73, 1-9.
- Fan, S., & Meyer, A. (2014). Evolution of genomic structural variation and genomic architecture in the adaptive radiations of African cichlid fishes. *Frontiers in Genetics*, 5, 163.
- Fang, B., Kemppainen, P., Momigliano, P., Feng, X., & Merilä, J. (2020). On the causes of geographically heterogeneous parallel evolution in sticklebacks. *Nature Ecology & Evolution*, 4(8), 1105-1115.
- Fang, S., Yukilevich, R., Chen, Y., Turissini, D. A., Zeng, K., Boussy, I. A., & Wu, C. I. (2012). Incompatibility and competitive exclusion of genomic segments between sibling *Drosophila* species. *PLoS Genetics*, 8(6), e1002795.
- Farhat, M. R., Shapiro, B. J., Kieser, K. J., Sultana, R., Jacobson, K. R., Victor, T. C., ... & Kaur, D. (2013). Genomic analysis identifies targets of convergent positive selection in drug-resistant *Mycobacterium tuberculosis*. *Nature Genetics*, 45, 1183-1189.
- Fay, J. C., & Wu, C. I. (2000). Hitchhiking under positive Darwinian selection. *Genetics*, 155(3), 1405-1413.

- Feng, J., Chi, P., Blanpied, T. A., Xu, Y., Magarinos, A. M., Ferreira, A., ... & Augustine GJ 2002. Regulation of neurotransmitter release by synapsin III. *Journal of Neuroscience*, 22, 4372-4380.
- Ferchaud, A. L., Pedersen, S. H., Bekkevold, D., Jian, J., Niu, Y., & Hansen, M. M. (2014). A low-density SNP array for analyzing differential selection in freshwater and marine populations of threespine stickleback (*Gasterosteus aculeatus*). *BMC Genomics*, 15(1), 867.
- Fernandez i Marti, A., & Dodd, R. S. (2018). Using CRISPR as a gene editing tool for validating adaptive gene function in tree landscape genomics. *Frontiers in Ecology and Evolution*, 6, 76.
- Ferrer-Admetlla, A., Liang, M., Korneliussen, T., & Nielsen, R. (2014). On detecting incomplete soft or hard selective sweeps using haplotype structure. *Molecular Biology and Evolution*, 31(5), 1275-1291.
- Feulner, P. G., Chain, F. J., Panchal, M., Eizaguirre, C., Kalbe, M., Lenz, T. L., ... & Reusch, T. B. (2013). Genome-wide patterns of standing genetic variation in a marine population of three-spined sticklebacks. *Molecular Ecology*, 22(3), 635-649.
- Filteau, M., Pavey, S. A., St-Cyr, J., Bernatchez, L. (2013). Gene coexpression networks reveal key drivers of phenotypic divergence in lake whitefish. *Molecular Biology and Evolution*, 30, 1384-1396.
- Finseth, F. R., Bondra, E., & Harrison, R. G. (2014). Selective constraint dominates the evolution of genes expressed in a novel reproductive gland. *Molecular Biology and Evolution*, 31, 3266-3281.
- Fisher, R. A. (1930). *The Genetical Theory of Natural Selection*. Oxford: Clarendon Press. 272 p.

- Foote, A. D., Liu, Y., Thomas, G. W., Vinař, T., Alföldi, J., Deng, J., ... & Khan, Z. (2015). Convergent evolution of the genomes of marine mammals. *Nature Genetics*, 47:272-275.
- Foote, A. D. (2018). Sympatric speciation in the genomic era. *Trends in Ecology & Evolution*, 33(2), 85-95.
- Fox, S. E., Christie, M. R., Marine, M., Priest, H. D., Mockler, T. C., & Blouin, M. S. (2014). Sequencing and characterization of the anadromous steelhead (*Oncorhynchus mykiss*) transcriptome. *Marine Genomics*, 15, 13-15.
- Fox, E. A., Wright, A. E., Fumagalli, M., & Vieira, F. G. (2019). ngsLD: evaluating linkage disequilibrium using genotype likelihoods. *Bioinformatics*, 35(19), 3855-3856.
- Franch-Gras, L., Hahn, C., García-Roger, E. M., Carmona, M. J., Serra, M., & Gómez, A. (2018). Genomic signatures of local adaptation to the degree of environmental predictability in rotifers. *Scientific Reports*, 8(1), 1-14.
- Franchini, P., Xiong, P., Fruciano, C., Schneider, R. F., Woltering, J. M., Hulsey, C. D., & Meyer, A. (2019). MicroRNA gene regulation in extremely young and parallel adaptive radiations of crater lake cichlid fish. *Molecular Biology and Evolution*, 36(11), 2498-2511.
- Fraser, D., Adams, C. E., & Huntingford, F. A. (1998). Trophic polymorphism among Arctic charr *Salvelinus alpinus* L., from Loch Ericht, Scotland. *Ecology of Freshwater Fish*, 7(4), 184-191.
- Friedman, S. T., & Adams, W. (1985). Estimation of gene flow into two seed orchards of loblolly pine (*Pinus taeda* L.). *Theoretical and Applied Genetics*, 69(5-6), 609-615.
- Fumagalli, M., Vieira, F. G., Linderoth, T., & Nielsen, R. (2014). ngsTools: methods for population genetics analyses from next-generation sequencing data. *Bioinformatics*, 30(10), 1486-1487.

- Gagnaire, P. A., Normandeau, E., Côté, C., Hansen, M. M., & Bernatchez, L. (2012). The genetic consequences of spatially varying selection in the panmictic American eel (*Anguilla rostrata*). *Genetics*, *190*, 725-736.
- Gagnaire, P. A., Pavey, S. A., Normandeau, E., & Bernatchez, L. (2013). The genetic architecture of reproductive isolation during speciation-with-gene-flow in lake whitefish species pairs assessed by RAD sequencing. *Evolution*, *67*(9), 2483-2497.
- Galinsky, K. J., Bhatia, G., Loh, P. R., Georgiev, S., Mukherjee, S., Patterson, N. J., & Price, A. L. (2016). Fast principal-component analysis reveals convergent evolution of ADH1B in Europe and East Asia. *The American Journal of Human Genetics*, *98*(3), 456-472.
- Garbarino, G., Costa, S., Pestarino, M., & Candiani, S. (2014). Differential expression of synapsin genes during early zebrafish development. *Neuroscience*, *280*, 351-367.
- Gardner, A. S., Walker, A. F., & Greer, R. B. (1988). Morphometric analysis of two ecologically distinct forms of Arctic charr, *Salvelinus alpinus* (L.), in Loch Rannoch, Scotland. *Journal of Fish Biology*, *32*(6), 901-910.
- Garud, N. R., Messer, P. W., Buzbas, E. O., & Petrov, D. A. (2015). Recent selective sweeps in North American *Drosophila melanogaster* show signatures of soft sweeps. *PLoS Genetics*, *11*(2).
- Garvin, M. R., Bielawski, J. P., & Gharrett, A. J. (2011). Positive Darwinian selection in the piston that powers proton pumps in complex I of the mitochondria of Pacific salmon. *PLoS One*, *6*(9), e24127.
- Gasman, S., Kalaidzidis, Y., & Zerial, M. (2003). RhoD regulates endosome dynamics through Diaphanous-related Formin and Src tyrosine kinase. *Nature Cell Biology*, *5*(3), 195-204.
- Gaudet, P., & Dessimoz, C. (2017). Gene ontology: pitfalls, biases, and remedies. In *The Gene Ontology Handbook* (pp. 189-205). Humana Press, New York, NY.

- Gavrilets, S. (2004). *Fitness landscapes and the origin of species*. Princeton University Press, Princeton, NJ.
- Germينو, M. J., Moser, A. M., & Sands, A. R. (2019). Adaptive variation, including local adaptation, requires decades to become evident in common gardens. *Ecological Applications*, 29(2), e01842.
- Gillard, G., Harvey, T. N., Gjuvsland, A., Jin, Y., Thomassen, M., Lien, S., ... & Sandve, S. R. (2018). Life-stage-associated remodelling of lipid metabolism regulation in Atlantic salmon. *Molecular Ecology*, 27(5), 1200-1213.
- Gish, W., & States, D. J. (1993). Identification of protein coding regions by database similarity search. *Nature Genetics*, 3(3), 266-272.
- Giska, I., Farelo, L., Pimenta, J., Seixas, F. A., Ferreira, M. S., Marques, J. P., ... & Magnussen, E. (2019). Introgression drives repeated evolution of winter coat color polymorphism in hares. *Proceedings of the National Academy of Sciences*, 116(48), 24150-24156.
- Gosset, C. C., & Bierne, N. (2013). Differential introgression from a sister species explains high FST outlier loci within a mussel species. *Journal of Evolutionary Biology*, 26(1), 14-26.
- Graur, D. (1985). Gene diversity in Hymenoptera. *Evolution*, 39(1), 190-199.
- Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., ... & Hansen, N. F. (2010). A draft sequence of the Neandertal genome. *Science*, 328(5979), 710-722.
- Griffiths, D. (2006). The direct contribution of fish to lake phosphorus cycles. *Ecology of Freshwater Fish*, 15, 86-95.
- Gudbrandsson, J., Ahi, E. P., Franzdottir, S. R., Kapralova, K. H., Kristjánsson, B. K., Steinhäuser, S. S., ... & Pálsson, A. (2015). The developmental transcriptome of contrasting Arctic charr (*Salvelinus alpinus*) morphs. *F1000Research*, 4, 136.

- Haldane, J. B. S. (1924). A mathematical theory of natural and artificial selection. Part II
The influence of partial self-fertilisation, inbreeding, assortative mating, and
selective fertilisation on the composition of Mendelian populations, and on natural
selection. *Biological Reviews*, 1(3), 158-163.
- Haldane, J. B. S. (1927). A mathematical theory of natural and artificial selection, part V:
selection and mutation. In *Mathematical Proceedings of the Cambridge
Philosophical Society* (Vol. 23, No. 7, pp. 838-844). Cambridge University Press.
- Haller, B. C. (2016). Eidos: a simple scripting language. URL:
http://benhaller.com/slim/Eidos_Manual.pdf.
- Haller, B. C., & Messer, P. W. (2017). SLiM 2: Flexible, interactive forward genetic
simulations. *Molecular Biology and Evolution*, 34(1), 230-240.
- Hämälä, T., & Savolainen, O. (2019). Genomic patterns of local adaptation under gene
flow in *Arabidopsis lyrata*. *Molecular Biology and Evolution*, 36(11), 2557-2571.
- Härer, A., Meyer, A., & Torres-Dowdall, J. (2018). Convergent phenotypic evolution of
the visual system via different molecular routes: How Neotropical cichlid fishes
adapt to novel light environments. *Evolution Letters*, 2(4), 341-354.
- Harris, A. M., Garud, N. R., & DeGiorgio, M. (2018). Detection and classification of hard
and soft sweeps from unphased genotypes by multilocus genotype identity.
Genetics, 210(4), 1429-1452.
- Hatfield, T., & Schluter, D. (1999). Ecological speciation in sticklebacks: environment-
dependent hybrid fitness. *Evolution*, 53, 866-873.
- He, Z., Xu, B., Buxbaum, J., & Ionita-Laza, I. (2019). A genome-wide scan statistic
framework for whole-genome sequence data analysis. *Nature Communications*,
10(1), 1-11.
- Hermisson, J., & Pennings, P. S. (2005). Soft sweeps: molecular population genetics of
adaptation from standing genetic variation. *Genetics*, 169(4), 2335-2352.

- Higgins, D. G., & Sharp, P. M. (1988). CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. *Gene*, 73(1), 237-244.
- Hill, M. O. 1973. Diversity and evenness: A unifying notation and its consequences. *Ecology*, 54, 427-432.
- Hodgins, K. A., Bock, D. G., Hahn, M. A., Heredia, S. M., Turner, K. G., & Rieseberg, L. H. (2015). Comparative genomics in the Asteraceae reveals little evidence for parallel evolutionary change in invasive taxa. *Molecular Ecology*, 24, 2226-2240.
- Hoegg, S., Brinkmann, H., Taylor, J. S., & Meyer, A. (2004). Phylogenetic timing of the fish-specific genome duplication correlates with the diversification of teleost fish. *Journal of Molecular Evolution*, 59, 190-203.
- Hohenlohe, P. A., & Magalhaes, I. S. (2020). The population genomics of parallel adaptation: lessons from threespine stickleback. *Population Genomics: Marine Organisms*, 249-276.
- Hohenlohe, P. A., Phillips, P. C., & Cresko, W. A. (2010). Using population genomics to detect selection in natural populations: key concepts and methodological considerations. *International Journal of Plant Sciences*, 171(9), 1059-1071.
- Holsinger, K. E., & Weir, B. S. (2009). Genetics in geographically structured populations: defining, estimating and interpreting F_{ST} . *Nature Reviews Genetics*, 10(9), 639-650.
- Hooker, O. E., Barry, J., Van Leeuwen, T. E., Lyle, A., Newton, J., Cunningham, P., & Adams, C. E. (2016). Morphological, ecological and behavioural differentiation of sympatric profundal and pelagic Arctic charr (*Salvelinus alpinus*) in Loch Dughail Scotland. *Hydrobiologia*, 783(1), 209-221.
- Houle, D., Hughes, K. A., Hoffmaster, D. K., Ihara, J., Assimacopoulos, S., & Charlesworth, B. (1994). The effects of spontaneous mutation on quantitative traits. I. Variances and covariances of life history traits. *Genetics*, 138, 773-785.

- Huang, X., Chen, X. G., & Armbruster, P. A. (2016). Comparative performance of transcriptome assembly methods for non-model organisms. *BMC Genomics*, *17*(1), 523.
- Huber, C. D., DeGiorgio, M., Hellmann, I., & Nielsen, R. (2016). Detecting recent selective sweeps while controlling for mutation rate and background selection. *Molecular Ecology*, *25*(1), 142-156.
- Hudson, A. G., Lundsgaard-Hansen, B., Lucek, K., Vonlanthen, P., & Seehausen, O. (2017). Managing cryptic biodiversity: Fine-scale intralacustrine speciation along a benthic gradient in Alpine whitefish (*Coregonus* spp.). *Evolutionary Applications*, *10*, 251-266.
- Hudson, M. E. (2008). Sequencing breakthroughs for genomic ecology and evolutionary biology. *Molecular Ecology Resources*, *8*(1), 3-17.
- Hudson, R. R. (2002). Generating samples under a Wright–Fisher neutral model of genetic variation. *Bioinformatics*, *18*(2), 337-338.
- Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernández-Plaza, A., Forslund, S. K., Cook, H., ... & von Mering, C. (2019). eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Research*, *47*(D1), D309-D314.
- Hufford, M. B., Lubinsky, P., Pyhäjärvi, T., Devengenzo, M. T., Ellstrand, N. C., & Ross-Ibarra, J. (2013). The genomic signature of crop-wild introgression in maize. *PLoS Genetics*, *9*(5).
- Hunt, B. G., Ometto, L., Wurm, Y., Shoemaker, D., Soojin, V. Y., Keller, L., & Goodisman, M. A. (2011). Relaxed selection is a precursor to the evolution of phenotypic plasticity. *Proceedings of the National Academy of Sciences*, *108*, 15936-15941.
- Innan, H., & Kondrashov, F. (2010). The evolution of gene duplications: classifying and distinguishing between models. *Nature Reviews Genetics*, *11*, 97.

- Jacobs, A., & Elmer, K. R. (2020). Alternative splicing and gene expression play contrasting roles in the parallel phenotypic evolution of a salmonid fish. *bioRxiv*, DOI: <https://doi.org/10.1101/2020.05.11.087973>.
- Jacobs, A., Carruthers, M., Eckmann, R., Yohannes, E., Adams, C. E., Behrmann-Godel, J., & Elmer, K. R. (2019). Rapid niche expansion by selection on functional genomic variation after ecosystem recovery. *Nature Ecology & Evolution*, 3(1), 77-86.
- Jacobs, A., Carruthers, M., Yurchenko, A., Gordeeva, N. V., Alekseyev, S. S., Hooker, O., ... & Adams, C. E., Elmer, K. R. (2020). Parallelism in eco-morphology and gene expression despite variable evolutionary and genomic backgrounds in a Holarctic fish. *PLoS Genetics*, 16(4), e1008658.
- Jacobs, A., Hughes, M. R., Robinson, P. C., Adams, C. E., & Elmer, K. R. (2018). The genetic architecture underlying the evolution of a rare piscivorous life history form in brown trout after secondary contact and strong introgression. *Genes*, 9(6), 280.
- Jacobs, A., Womack, R., Chen, M., Gharbi, K., & Elmer, K. R. (2017). Significant synteny and colocalization of ecologically relevant quantitative trait loci within and across species of salmonid fishes. *Genetics*, 207(2), 741-754.
- Jensen, J. D., Thornton, K. R., & Andolfatto, P. (2008). An approximate Bayesian estimator suggests strong, recurrent selective sweeps in *Drosophila*. *PLoS Genetics*, 4(9).
- Jensen, L. J., Julien, P., Kuhn, M., von Mering, C., Muller, J., Doerks, T., & Bork, P. (2007). eggNOG: automated construction and annotation of orthologous groups of genes. *Nucleic Acids Research*, 36(suppl_1), D250-D254.
- Jeukens, J., Renaut, S., St-Cyr, J., Nolte, A. W., & Bernatchez, L. (2010). The transcriptomics of sympatric dwarf and normal lake whitefish (*Coregonus clupeaformis* spp., Salmonidae) divergence as revealed by next-generation sequencing. *Molecular Ecology*, 19, 5389-5403.

- Jobling, M., Johansen, S. J. S., Foshaug, H., Burkow, I. C., & Jørgensen, E. H. (1998). Lipid dynamics in anadromous Arctic charr, *Salvelinus alpinus* (L.): seasonal variations in lipid storage depots and lipid class composition. *Fish Physiology and Biochemistry*, 18(3), 225-240.
- Johannesson, K., & Butlin, R. K. (2017). What explains rare and conspicuous colours in a snail? A test of time-series data against models of drift, migration or selection. *Heredity*, 118(1), 21-30.
- Johnston, S. E., Orell, P., Pritchard, V. L., Kent, M. P., Lien, S., Niemelä, E., ... & Primmer, C. R. (2014). Genome-wide SNP analysis reveals a genetic basis for sea-age variation in a wild population of Atlantic salmon (*Salmo salar*). *Molecular Ecology*, 23, 3452-3468.
- Jones, P., Binns, D., Chang, H. Y., Fraser, M., Li, W., McAnulla, C., ... & Pesseat, S. (2014). InterProScan 5: genome-scale protein function classification. *Bioinformatics*, 30, 1236-1240.
- Jones, F. C., Grabherr, M. G., Chan, Y. F., Russell, P., Mauceli, E., Johnson, J., ... & Birney, E. (2012). The genomic basis of adaptive evolution in threespine sticklebacks. *Nature*, 484(7392), 55-61.
- Jones, M. R., Mills, L. S., Alves, P. C., Callahan, C. M., Alves, J. M., Lafferty, D. J., ... & Good, J. M. (2018). Adaptive introgression underlies polymorphic seasonal camouflage in snowshoe hares. *Science*, 360(6395), 1355-1358.
- Jonsson, B., & Jonsson, N. (2001). Polymorphism and speciation in Arctic charr. *Journal of Fish Biology*, 58, 605-638.
- Jordan, G., & Goldman, N. (2011). The effects of alignment error and alignment filtering on the sitewise detection of positive selection. *Molecular Biology and Evolution*, 29, 1125-1139.
- Jost, L. (2010). Independence of alpha and beta diversities. *Ecology*, 91(7), 1969-1974.

- Kao, H. T., Li, P., Chao, H. M., Janoschka, S., Pham, K., Feng, J., ... & Porton, B. (2008). Early involvement of synapsin III in neural progenitor cell development in the adult hippocampus. *Journal of Comparative Neurology*, 507, 1860-1870.
- Kapralova, K. H., Gudbrandsson, J., Reynisdottir, S., Santos, C. B., Baltanás, V. C., Maier, V. H., ... & Palsson, A. (2013). Differentiation at the MHCII α and Cath2 loci in sympatric *Salvelinus alpinus* resource morphs in Lake Thingvallavatn. *PLoS ONE*, 8, e69402.
- Kautt, A. F., Machado-Schiaffino, G., & Meyer, A. (2016). Multispecies outcomes of sympatric speciation after admixture with the source population in two radiations of Nicaraguan crater lake cichlids. *PLoS Genetics*, 12(6).
- Kavanagh, K. D., Haugen, T. O., Gregersen, F., Jernvall, J., & Vøllestad, L. A. (2010). Contemporary temperature-driven divergence in a Nordic freshwater fish under conditions commonly thought to hinder adaptation. *BMC Evolutionary Biology*, 10(1), 350.
- Kawakami, T., Smeds, L., Backström, N., Husby, A., Qvarnström, A., Mugal, C. F., ... & Ellegren, H. (2014). A high-density linkage map enables a second-generation collared flycatcher genome assembly and reveals the patterns of avian recombination rate variation and chromosomal evolution. *Molecular Ecology*, 23(16), 4035-4058.
- Keefer, M. L., & Caudill, C. C. (2014). Homing and straying by anadromous salmonids: a review of mechanisms and rates. *Reviews in Fish Biology and Fisheries*, 24(1), 333-368.
- Kelly, J. K. (1997). A test of neutrality based on interlocus associations. *Genetics*, 146(3), 1197-1206.
- Kenney, A. M., & Sweigart, A. L. (2016). Reproductive isolation and introgression between sympatric *Mimulus* species. *Molecular Ecology*, 25(11), 2499-2517.
- Kern, A. D., & Hahn, M. W. (2018). The neutral theory in light of natural selection. *Molecular Biology and Evolution*, 35(6), 1366-1371.

- Kern, A. D., & Schrider, D. R. (2018). diploS/HIC: an updated approach to classifying selective sweeps. *G3: Genes, Genomes, Genetics*, 8(6), 1959-1970.
- Kim, Y., & Nielsen, R. (2004). Linkage disequilibrium as a signature of selective sweeps. *Genetics*, 167(3), 1513-1524.
- Kimura, M. (1961). Natural selection as the process of accumulating genetic information in adaptive evolution. *Genetics Research*, 2(1), 127-140.
- Kimura, M., & Ohta, T. (1971). Protein polymorphism as a phase of molecular evolution. *Nature*, 229(5285), 467-469.
- Kirch, M., Romundset, A., Gilbert, M. T. P., Jones, F. C., & Foote, A. D. (2020). Pleistocene stickleback genomes reveal the constraints on parallel evolution. *bioRxiv*, DOI: <https://doi.org/10.1101/2020.08.12.248427>.
- Kirubakaran, T. G., Grove, H., Kent, M. P., Sandve, S. R., Baranski, M., Nome, T., ... & Sonesson, A. (2016). Two adjacent inversions maintain genomic differentiation between migratory and stationary ecotypes of Atlantic cod. *Molecular Ecology*, 25(10), 2130-2143.
- Klemetsen, A., Amundsen, P. A., Dempson, J. B., Jonsson, B., Jonsson, N., O'Connell, M. F., & Mortensen, E. (2003). Atlantic salmon *Salmo salar* L., brown trout *Salmo trutta* L. and Arctic charr *Salvelinus alpinus* (L.): a review of aspects of their life histories. *Ecology of Freshwater Fish*, 12, 1-59.
- Klemetsen, A. (2010). The charr problem revisited: exceptional phenotypic plasticity promotes ecological speciation in postglacial lakes. *Freshwater Reviews*, 3(1), 49-74.
- Klemetsen, A. (2013). The most variable vertebrate on Earth. *Journal of Ichthyology*, 53(10), 781-791.
- Knudsen, R., Klemetsen, A., Alekseyev, S., Adams, C. E., & Power, M. (2016). The role of *Salvelinus* in contemporary studies of evolution, trophic ecology and anthropogenic change. *Hydrobiologia*, 783(1), 1-9.

- Koblmüller, S., Odhiambo, E. A., Sinyinza, D., Sturmbauer, C., & Sefc, K. M. (2015). Big fish, little divergence: phylogeography of Lake Tanganyika's giant cichlid, *Boulengerochromis microlepis*. *Hydrobiologia*, 748, 29-38.
- Koepfli, K. P., Paten, B., Genome 10K Community of Scientists, & O'Brien, S. J. (2015). The Genome 10K Project: a way forward. *Annu. Rev. Anim. Biosci.*, 3(1), 57-111.
- Kondrashov, F. A., Rogozin, I. B., Wolf, Y. I., & Koonin, E. V. (2002). Selection in the evolution of gene duplications. *Genome Biology*, 3, 8.1-8.9.
- Koonin, E. V., & Mushegian, A. R. (1996). Complete genome sequences of cellular life forms: glimpses of theoretical evolutionary genomics. *Current Opinion in Genetics & Development*, 6(6), 757-762.
- Korneliussen, T. S., Albrechtsen, A., & Nielsen, R. (2014). ANGSD: analysis of next generation sequencing data. *BMC Bioinformatics*, 15(1), 356.
- Kottelat, M., Freyhof, J. (2007). Handbook of European freshwater fishes. Publications Kottelat, Cornol and Freyhof, Berlin.
- Kowalko, J. E., Rohner, N., Linden, T. A., Rompani, S. B., Warren, W. C., Borowsky, R., ... & Yoshizawa, M. (2013). Convergence in feeding posture occurs through different genetic loci in independently evolved cave populations of *Astyanax mexicanus*. *Proceedings of the National Academy of Sciences*, 110(42), 16933-16938.
- Kristjánsson, B. K., Skúlason, S., Snorrason, S. S., & Noakes, D. L. (2012). Fine-scale parallel patterns in diversity of small benthic Arctic charr (*Salvelinus alpinus*) in relation to the ecology of lava/groundwater habitats. *Ecology and Evolution*, 2(6), 1099-1112.
- Kulmuni, J., Butlin, R. K., Lucek, K., Savolainen, V., & Westram, A. M. (2020). Towards the completion of speciation: the evolution of reproductive isolation beyond the first barriers. *Proceedings of the Royal Society B*, 375(1806), 20190528.

- Kumar, S., Stecher, G., Li, M., Knyaz, C., & Tamura, K. (2018). MEGA X: molecular evolutionary genetics analysis across computing platforms. *Molecular Biology and Evolution*, 35(6), 1547-1549.
- Kumar, S., Tomura, K., & Nei, M. (1993). MEGA: Molecular evolution genetics analysis, version 1.0. University Park (PA): Pennsylvania State University.
- Kurko, J., Debes, P. V., House, A. H., Aykanat, T., Erkinaro, J., & Primmer, C. R. (2020). Transcription profiles of age-at-maturity-associated genes suggest cell fate commitment regulation as a key factor in the Atlantic salmon maturation process. *G3: Genes, Genomes, Genetics*, 10(1), 235-246.
- Laayouni, H., Oosting, M., Luisi, P., Ioana, M., Alonso, S., Ricaño-Ponce, I., ... & van der Meer, J. W. (2014). Convergent evolution in European and Roma populations reveals pressure exerted by plague on Toll-like receptors. *Proceedings of the National Academy of Sciences*, 111, 2668-2673.
- Lamichhaney, S., Fan, G., Widemo, F., Gunnarsson, U., Thalmann, D. S., Hoepfner, M. P., ... & Chen, W. (2016). Structural genomic changes underlie alternative reproductive strategies in the ruff (*Philomachus pugnax*). *Nature Genetics*, 48, 84.
- Lamichhaney, S., Han, F., Webster, M. T., Andersson, L., Grant, B. R., & Grant, P. R. (2018). Rapid hybrid speciation in Darwin's finches. *Science*, 359(6372), 224-228.
- Langerhans, R. B. (2018). Predictability and parallelism of multitrait adaptation. *Journal of Heredity*, 109(1), 59-70.
- Laporte, M., Dalziel, A. C., Martin, N., & Bernatchez, L. (2016). Adaptation and acclimation of traits associated with swimming capacity in Lake Whitefish (*Coregonus clupeaformis*) ecotypes. *BMC Evolutionary Biology*, 16, 160.
- Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., ... & Thompson, J. D. (2007). Clustal W and Clustal X version 2.0. *Bioinformatics*, 23(21), 2947-2948.

- Larson, W. A., Dann, T. H., Limborg, M. T., McKinney, G. J., Seeb, J. E., & Seeb, L. W. (2019). Parallel signatures of selection at genomic islands of divergence and the MHC in ecotypes of sockeye salmon across Alaska. *Molecular Ecology*, DOI: 10.1111/mec.15082.
- Larson, W. A., Limborg, M. T., McKinney, G. J., Schindler, D. E., Seeb, J. E., & Seeb, L. W. (2017). Genomic islands of divergence linked to ecotypic variation in sockeye salmon. *Molecular Ecology*, 26(2), 554-570.
- Laurent, B., Moinard, M., Spataro, C., Ponts, N., Barreau, C., & Foulongne-Oriol, M. (2017). Landscape of genomic diversity and host adaptation in *Fusarium graminearum*. *BMC Genomics*, 18(1), 203.
- Laurentino, T. G., Moser, D., Roesti, M., Ammann, M., Frey, A., Ronco, F., ... & Berner, D. (2020). Genomic release-recapture experiment in the wild reveals within-generation polygenic selection in stickleback fish. *Nature Communications*, 11(1), 1-9.
- Lawlor, D. A., Zemmour, J., Ennis, P. D., & Parham, P. (1990). Evolution of class-I MHC genes and proteins: from natural selection to thymic selection. *Annual Review of Immunology*, 8(1), 23-63.
- Lawrence, M., Huber, W., Pages, H., Aboyoun, P., Carlson, M., Gentleman, R., ... & Carey, V. J. (2013). Software for computing and annotating genomic ranges. *PLoS Computational Biology*, 9(8), e1003118.
- Lawson, L. P., & Petren, K. (2017). The adaptive genomic landscape of beak morphology in Darwin's finches. *Molecular Ecology*, 26(19), 4978-4989.
- Lecaudey, L. A., Schliwen, U. K., Osinov, A. G., Taylor, E. B., Bernatchez, L., & Weiss, S. J. (2018). Inferring phylogenetic structure, hybridization and divergence times within Salmoninae (Teleostei: Salmonidae) using RAD-sequencing. *Molecular Phylogenetics and Evolution*, 124, 82-99.

- Lee, J., Freddolino, P. L., & Zhang, Y. (2017). Ab initio protein structure prediction. In *From Protein Structure to Function with Bioinformatics* (pp. 3-35). Springer, Dordrecht.
- Lee, S. (2019). The association of genetically controlled CpG methylation (cg158269415) of protein tyrosine phosphatase, receptor type N2 (PTPRN2) with childhood obesity. *Scientific Reports*, 9(1), 1-7.
- Lefort, V., Desper, R., & Gascuel, O. (2015). FastME 2.0: a comprehensive, accurate, and fast distance-based phylogeny inference program. *Molecular Biology and Evolution*, 32, 2798-2800.
- Lehnert, S. J., Kess, T., Bentzen, P., Kent, M. P., Lien, S., Gilbey, J., ... & Bradbury, I. R. (2019). Genomic signatures and correlates of widespread population declines in salmon. *Nature Communications*, 10(1), 1-10.
- Lemopoulos, A., Uusi-Heikkilä, S., Huusko, A., Vasemägi, A., & Vainikka, A. (2018). Comparison of migratory and resident populations of brown trout reveals candidate genes for migration tendency. *Genome Biology and Evolution*, 10(6), 1493-1503.
- Lenormand, T. (2002). Gene flow and the limits to natural selection. *Trends in Ecology & Evolution*, 17(4), 183-189.
- Lenski, R. E. (2017). Experimental evolution and the dynamics of adaptation and genome evolution in microbial populations. *The ISME journal*, 11(10), 2181-2194.
- Lepage, S. E., & Bruce, A. E. (2010). Zebrafish epiboly: mechanics and mechanisms. *International Journal of Developmental Biology*, 54(8-9), 1213-1228.
- Li, H., & Durbin, R. (2010). Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics*, 26(5), 589-595.
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., ... & Durbin, R. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, 25(16), 2078-2079.

- Liang, Y., Meyer, A., & Kratochwil, C. F. (2020). Neural innervation as a potential trigger of morphological color change and sexual dimorphism in cichlid fish. *Scientific Reports*, 10(1), 1-16.
- Liao, Y., Wang, J., Jaehnig, E. J., Shi, Z., & Zhang, B. (2019). WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. *Nucleic Acids Research*, 47(W1), W199-W205.
- Lien, S., Koop, B. F., Sandve, S. R., Miller, J. R., Kent, M. P., Nome, T., ... & Grammes, F. (2016). The Atlantic salmon genome provides insights into rediploidization. *Nature*, 533(7602), 200-205.
- Lindtke, D., & Buerkle, C. A. (2015). The genetic architecture of hybrid incompatibilities and their effect on barriers to introgression in secondary contact. *Evolution*, 69(8), 1987-2004.
- Liu, H., Fu, B., Pang, M., Feng, X., Wang, X., Yu, X., & Tong, J. (2016). QTL fine mapping and identification of candidate genes for growth-related traits in bighead carp (*Hypophthalmichthys nobilis*). *Aquaculture*, 465, 134-143.
- Lobkovsky, A. E., & Koonin, E. V. (2012). Replaying the tape of life: quantification of the predictability of evolution. *Frontiers in Genetics*, 3, 246.
- Lohmueller, K. E. (2014). The impact of population demography and selection on the genetic architecture of complex traits. *PLoS Genetics*, 10(5), e1004379.
- Long, A., Liti, G., Luptak, A., & Tenaillon, O. (2015). Elucidating the molecular architecture of adaptation via evolve and resequence experiments. *Nature Reviews Genetics*, 16(10), 567.
- López-Fernández, H., Arbour, J. H., Winemiller, K., & Honeycutt, R. L. (2013). Testing for ancient adaptive radiations in Neotropical cichlid fishes. *Evolution*, 67, 1321-1337.

- Lopez, M., Choin, J., Sikora, M., Siddle, K., Harmant, C., Costa, H. A., ... & Le Bomin, S. (2019). Genomic evidence for local adaptation of hunter-gatherers to the African rainforest. *Current Biology*, 29(17), 2926-2935.
- Lotterhos, K. E. (2019). The effect of neutral recombination variation on genome scans for selection. *G3: Genes, Genomes, Genetics*, 9(6), 1851-1867.
- Lotterhos, K. E., & Whitlock, M. C. (2014). Evaluation of demographic history and neutral parameterization on the performance of FST outlier tests. *Molecular Ecology*, 23(9), 2178-2192.
- Lotterhos, K. E., Card, D. C., Schaal, S. M., Wang, L., Collins, C., & Verity, B. (2017). Composite measures of selection can improve the signal-to-noise ratio in genome scans. *Methods in Ecology and Evolution*, 8(6), 717-727.
- Lowenstein, J. M. (1986). Molecular phylogenetics. *Annual Review of Earth and Planetary Sciences*, 14(1), 71-83.
- Lowry, D. B., Hoban, S., Kelley, J. L., Lotterhos, K. E., Reed, L. K., Antolin, M. F., & Storfer, A. (2017). Breaking RAD: An evaluation of the utility of restriction site-associated DNA sequencing for genome scans of adaptation. *Molecular Ecology Resources*, 17(2), 142-152.
- Löytynoja, A. (2014). Phylogeny-aware alignment with PRANK. In *Multiple Sequence Alignment Methods* (pp. 155-170). Humana Press, Totowa, NJ.
- Lu, G., & Bernatchez, L. (1999). Correlated trophic specialization and genetic divergence in sympatric lake whitefish ecotypes (*Coregonus clupeaformis*): support for the ecological speciation hypothesis. *Evolution*, 53, 1491-1505.
- Lucek, K., Keller, I., Nolte, A. W., & Seehausen, O. (2018). Distinct colonization waves underlie the diversification of the freshwater sculpin (*Cottus gobio*) in the Central European Alpine region. *Journal of Evolutionary Biology*, 31(9), 1254-1267.

- Lundsgaard-Hansen, B., Matthews, B., Vonlanthen, P., Taverna, A., & Seehausen, O. (2013). Adaptive plasticity and genetic divergence in feeding efficiency during parallel adaptive radiation of whitefish (*Coregonus* spp.). *Journal of Evolutionary Biology*, 26(3), 483-498.
- Lynch, M., & Conery, J. S. (2000). The evolutionary fate and consequences of duplicate genes. *Science*, 290, 1151-1155.
- Lynch, M., Ackerman, M. S., Gout, J. F., Long, H., Sung, W., Thomas, W. K., & Foster, P. L. (2016). Genetic drift, selection and the evolution of the mutation rate. *Nature Reviews Genetics*, 17(11), 704.
- Mack, K. L., & Nachman, M. W. (2017). Gene regulation and speciation. *Trends in Genetics*, 33(1), 68-80.
- Mack, K. L., Campbell, P., & Nachman, M. W. (2016). Gene regulation and speciation in house mice. *Genome Research*, 26(4), 451-461.
- Macqueen, D. J., & Johnston, I. A. (2014). A well-constrained estimate for the timing of the salmonid whole genome duplication reveals major decoupling from species diversification. *Proc. R. Soc. B*, 281, 20132881.
- Macqueen, D. J., Primmer, C. R., Houston, R. D. et al. (2017). Functional Annotation of All Salmonid Genomes (FAASG): an international initiative supporting future salmonid research, conservation and aquaculture. *BMC Genomics*, 18, 484.
- Mahler, D. L., Ingram, T., Revell, L. J., & Losos, J. B. (2013). Exceptional convergence on the macroevolutionary landscape in island lizard radiations. *Science*, 341(6143), 292-295.
- Mahler, D. L., Revell, L. J., Glor, R. E., & Losos, J. B. (2010). Ecological opportunity and the rate of morphological evolution in the diversification of Greater Antillean anoles. *Evolution*, 64(9), 2731-2745.

- Mahmoud, M., Gobet, N., Cruz-Dávalos, D. I., Mounier, N., Dessimoz, C., & Sedlazeck, F. J. (2019). Structural variant calling: the long and the short of it. *Genome Biology*, 20(1), 246.
- Malinsky, M., Challis, R. J., Tyers, A. M., Schiffels, S., Terai, Y., Ngatunga, B. P., ... & Turner, G. F. (2015). Genomic islands of speciation separate cichlid ecomorphs in an East African crater lake. *Science*, 350(6267), 1493-1498.
- Malinsky, M., Svoldal, H., Tyers, A. M., Miska, E. A., Genner, M. J., Turner, G. F., & Durbin, R. (2018). Whole-genome sequences of Malawi cichlids reveal multiple radiations interconnected by gene flow. *Nature Ecology & Evolution*, 2(12), 1940-1955.
- Malmstrøm, M., Matschiner, M., Tørresen, O. K., Star, B., Snipen, L. G., Hansen, T. F., ... & Stenseth, N. C. (2016). Evolution of the immune system influences speciation rates in teleost fishes. *Nature Genetics*, 48(10), 1204-1210.
- Manceau, M., Domingues, V. S., Linnen, C. R., Rosenblum, E. B., & Hoekstra, H. E. (2010). Convergence in pigmentation at multiple levels: mutations, genes and function. *Phil. Trans. R. Soc. B*, 365:2439-2450.
- Markova-Raina, P., & Petrov, D. (2011). High sensitivity to aligner and high rate of false positives in the estimates of positive selection in the 12 *Drosophila* genomes. *Genome Research*, 21, 863-874.
- Marques, D. A., Jones, F. C., Di Palma, F., Kingsley, D. M., & Reimchen, T. E. (2018). Experimental evidence for rapid genomic adaptation to a new niche in an adaptive radiation. *Nature Ecology & Evolution*, 2(7), 1128-1138.
- Marques, D. A., Lucek, K., Sousa, V. C., Excoffier, L., & Seehausen, O. (2019). Admixture between old lineages facilitated contemporary ecological speciation in Lake Constance stickleback. *Nature Communications*, 10(1), 1-14.
- Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, 17(1), 10-12.

- Martin, S. H., Davey, J. W., Salazar, C., & Jiggins, C. D. (2019). Recombination rate variation shapes barriers to introgression across butterfly genomes. *PLoS Biology*, *17*(2), e2006288.
- Matschiner, M., Musilová, Z., Barth, J. M., Starostová, Z., Salzburger, W., Steel, M., & Bouckaert, R. (2017). Bayesian phylogenetic estimation of clade ages supports trans-Atlantic dispersal of cichlid fishes. *Systematic Biology*, *66*, 3-22.
- Matthews, W. J. (1998). Patterns in freshwater fish ecology. *The University of Chicago Press*, New York.
- Matthews, B., Marchinko, K. B., Bolnick, D. I., & Mazumder, A. (2010). Specialization of trophic position and habitat use by sticklebacks in an adaptive radiation. *Ecology*, *91*(4), 1025-1034.
- Matthey-Doret, R., & Whitlock, M. C. (2019). Background selection and FST: consequences for detecting local adaptation. *Molecular Ecology*, *28*(17), 3902-3914.
- McGee, M. D., Borstein, S. R., Meier, J. I., Marques, D. A., Mwaiko, S., Taabu, A., ... & Seehausen, O. (2020). The ecological and genomic basis of explosive adaptive radiation. *Nature*, *586*(7827), 75-79.
- McGirr, J. A., & Martin, C. H. (2018). Parallel evolution of gene expression between trophic specialists despite divergent genotypes and morphologies. *Evolution Letters*, *2*(2), 62-75.
- McGirr, J. A., & Martin, C. H. (2020). Ecological divergence in sympatry causes gene misexpression in hybrids. *Molecular Ecology*, *29*(14), 2707-2721.
- McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., ... & DePristo, M. A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297-1303.

- McPhee, M. V., Utter, F., Stanford, J. A., Kuzishchin, K. V., Savvaitova, K. A., Pavlov, D. S., & Allendorf, F. W. (2007). Population structure and partial anadromy in *Oncorhynchus mykiss* from Kamchatka: relevance for conservation strategies around the Pacific Rim. *Ecology of Freshwater Fish*, 16, 539-547.
- Meier, J. I., Marques, D. A., Wagner, C. E., Excoffier, L., & Seehausen, O. (2018). Genomics of parallel ecological speciation in Lake Victoria cichlids. *Molecular Biology and Evolution*, 35(6), 1489-1506.
- Meisner, J., & Albrechtsen, A. (2018). Inferring population structure and admixture proportions in low-depth NGS data. *Genetics*, 210(2), 719-731.
- Messer, P. W. (2013). SLiM: simulating evolution with selection and linkage. *Genetics*, 194(4), 1037-1039.
- Messer, P. W., & Petrov, D. A. (2013). Population genomics of rapid adaptation by soft selective sweeps. *Trends in Ecology & Evolution*, 28(11), 659-669.
- Mikkelsen, T. S., Wakefield, M. J., Aken, B., Amemiya, C. T., Chang, J. L., Duke, S., ... & Jurka, J. (2007). Genome of the marsupial *Monodelphis domestica* reveals innovation in non-coding sequences. *Nature*, 447(7141), 167-177.
- Miller, S. E., Roesti, M., & Schluter, D. (2019). A single interacting species leads to widespread parallel evolution of the stickleback genome. *Current Biology*, 29(3), 530-537.
- Mitchell, S., Reeve, R., & White, T. J. (2020). rdiversity. R package version 2.0. <https://github.com/boydorr/rdiversity>
- Mitterboeck, T. F., Liu, S., Adamowicz, S. J., Fu, J., Zhang, R., Song, W., ... & Zhou, X. (2017). Positive and relaxed selection associated with flight evolution and loss in insect transcriptomes. *GigaScience*, 6, 1-14.
- Moeller, D. A., Tenailon, M. I., & Tiffin, P. (2007). Population structure and its effects on patterns of nucleotide polymorphism in teosinte (*Zea mays* ssp. *parviglumis*). *Genetics*, 176(3), 1799-1809.

- Moest, M., Van Belleghem, S. M., James, J. E., Salazar, C., Martin, S. H., Barker, S. L., ... & Steiner, F. M. (2020). Selective sweeps on novel and introgressed variation shape mimicry loci in a butterfly adaptive radiation. *PLoS Biology*, *18*(2), e3000597.
- Moore, J. S., Harris, L. N., Le Luyer, J., Sutherland, B. J., Rougemont, Q., Tallman, R. F., ... & Bernatchez, L. (2017). Genomics and telemetry suggest a role for migration harshness in determining overwintering habitat choice, but not gene flow, in anadromous Arctic Char. *Molecular Ecology*, *26*, 6784-6800.
- Morales, H. E., Faria, R., Johannesson, K., Larsson, T., Panova, M., Westram, A. M., & Butlin, R. K. (2019). Genomic architecture of parallel ecological divergence: beyond a single environmental contrast. *Science Advances*, *5*(12), eaav9963.
- Morandin, C., Tin, M. M., Abril, S., Gómez, C., Pontieri, L., Schiøtt, M., ... & Mikheyev, A. S. (2016). Comparative transcriptomics reveals the conserved building blocks involved in parallel evolution of diverse phenotypic traits in ants. *Genome Biology*, *17*, 43.
- Morgan, K., Harr, B., White, M. A., Payseur, B. A., & Turner, L. M. (2020). Disrupted gene networks in subfertile hybrid house mice. *Molecular Biology and Evolution*, *37*(6), 1547-1562.
- Morgulis, A., Coulouris, G., Raytselis, Y., Madden, T. L., Agarwala, R., & Schäffer, A. A. (2008). Database indexing for production MegaBLAST searches. *Bioinformatics*, *24*(16), 1757-1764.
- Muir, A. M., Hansen, M. J., Bronte, C. R., & Krueger, C. C. (2016). If Arctic charr *Salvelinus alpinus* is ‘the most diverse vertebrate’, what is the lake charr *Salvelinus namaycush*?. *Fish and Fisheries*, *17*, 1194-1207.
- Murrell, B., Weaver, S., Smith, M. D., Wertheim, J. O., Murrell, S., Aylward, A., ... & Scheffler, K. (2015). Gene-wide identification of episodic selection. *Molecular Biology and Evolution*, *32*, 1365-1371.

- Muschick, M., Barluenga, M., Salzburger, W., & Meyer, A. (2011). Adaptive phenotypic plasticity in the Midas cichlid fish pharyngeal jaw and its relevance in adaptive radiation. *BMC Evolutionary Biology*, 11(1), 116.
- Myhre, S., Tveit, H., Mollestad, T., & Lægreid, A. (2006). Additional gene ontology structure for improved biological reasoning. *Bioinformatics*, 22, 2020-2027.
- Naciri, Y., & Linder, H. P. (2020). The genetics of evolutionary radiations. *Biological Reviews*, 95(4), 1055-1072.
- Nadeau, N. J., Whibley, A., Jones, R. T., Davey, J. W., Dasmahapatra, K. K., Baxter, S. W., ... & Mallet, J. (2012). Genomic islands of divergence in hybridizing *Heliconius* butterflies identified by large-scale targeted sequencing. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1587), 343-353.
- Narum, S. R., & Hess, J. E. (2011). Comparison of FST outlier tests for SNP loci under selection. *Molecular Ecology Resources*, 11, 184-194.
- Natarajan, C., Projecto-Garcia, J., Moriyama, H., Weber, R. E., Muñoz-Fuentes, V., Green, A. J., ... & Smith, M. M. (2015). Convergent evolution of hemoglobin function in high-altitude Andean waterfowl involves limited parallelism at the molecular sequence level. *PLoS Genetics*, 11(12), e1005681.
- Nei, M., & Li, W. H. (1979). Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences*, 76(10), 5269-5273.
- Nelson, T. C., & Cresko, W. A. (2018). Ancient genomic variation underlies repeated ecological adaptation in young stickleback populations. *Evolution Letters*, 2(1), 9-21.
- Ng, P. C., & Henikoff, S. (2006). Predicting the effects of amino acid substitutions on protein function. *Annu. Rev. Genomics Hum. Genet.*, 7, 61-80.
- Nielsen, R., Paul, J. S., Albrechtsen, A., & Song, Y. S. (2011). Genotype and SNP calling from next-generation sequencing data. *Nature Reviews Genetics*, 12(6), 443-451.

- Nielsen, R., Williamson, S., Kim, Y., Hubisz, M. J., Clark, A. G., & Bustamante, C. (2005). Genomic scans for selective sweeps using SNP data. *Genome Research*, 15(11), 1566-1575.
- Nolte, A. W., Freyhof, J., & Tautz, D. (2006). When invaders meet locally adapted types: rapid moulding of hybrid zones between sculpins (*Cottus*, Pisces) in the Rhine system. *Molecular Ecology*, 15(7), 1983-1993.
- Noor, M. A. (1999). Reinforcement and other consequences of sympatry. *Heredity*, 83(5), 503-508.
- Norman, J. D., Ferguson, M. M., & Danzmann, R. G. (2014). Transcriptomics of salinity tolerance capacity in Arctic charr (*Salvelinus alpinus*): a comparison of gene expression profiles between divergent QTL genotypes. *Physiological Genomics*, 46(4), 123-137.
- Nosil, P. (2008). Speciation with gene flow could be common. *Molecular Ecology*, 17(9), 2103-2106.
- Nosil, P. (2012). *Ecological speciation*. Oxford University Press, Oxford, UK.
- Nosil, P., Villoutreix, R., de Carvalho, C. F., Farkas, T. E., Soria-Carrasco, V., Feder, J. L., ... & Gompert, Z. (2018). Natural selection and the predictability of evolution in *Timema* stick insects. *Science*, 359(6377), 765-770.
- Nozawa, M., Suzuki, Y., & Nei, M. (2009). Reliabilities of identifying positive selection by the branch-site and the site-prediction methods. *Proceedings of the National Academy of Sciences*, 106, 6700-6705.
- O'Connor, E. A., Strandh, M., Hasselquist, D., Nilsson, J. Å., & Westerdahl, H. (2016). The evolution of highly variable immunity genes across a passerine bird radiation. *Molecular Ecology*, 25(4), 977-989.
- Oke, K. B., Rolshausen, G., LeBlond, C., & Hendry, A. P. (2017). How parallel is parallel evolution? A comparative analysis in fishes. *The American Naturalist*, 190(1), 1-16.

- Orr, H. A., Masly, J. P., & Presgraves, D. C. (2004). *Speciation genes. Current Opinion in Genetics & Development, 14*, 675-679.
- Osborne, O. G., Ciezarek, A., Wilson, T., Crayn, D., Hutton, I., Baker, W. J., ... & Savolainen, V. (2019). Speciation in *Howea* palms occurred in sympatry, was preceded by ancestral admixture, and was associated with edaphic and phenological adaptation. *Molecular Biology and Evolution, 36*(12), 2682-2697.
- Østbye, K., Amundsen, P. A., Bernatchez, L., Klemetsen, A., Knudsen, R., Kristoffersen, R., ... & Hindar, K. (2006). Parallel evolution of ecomorphological traits in the European whitefish *Coregonus lavaretus* (L.) species complex during postglacial times. *Molecular Ecology, 15*(13), 3983-4001.
- Pai, A. A., & Gilad, Y. (2014). Comparative studies of gene regulatory mechanisms. *Current Opinion in Genetics & Development, 29*, 68-74.
- Papadopoulos, A. S., Igea, J., Smith, T. P., Hutton, I., Baker, W. J., Butlin, R. K., & Savolainen, V. (2019). Ecological speciation in sympatric palms: 4. Demographic analyses support speciation of *Howea* in the face of high gene flow. *Evolution, 73*(9), 1996-2002.
- Paradis, E., Claude, J., & Strimmer, K. (2004). APE: analyses of phylogenetics and evolution in R language. *Bioinformatics, 20*, 289-290.
- Pardo-Diaz, C., Salazar, C., Baxter, S. W., Merot, C., Figueiredo-Ready, W., Joron, M., ... & Jiggins, C. D. (2012). Adaptive introgression across species boundaries in *Heliconius* butterflies. *PLoS Genetics, 8*(6), e1002752.
- Pasquier, J., Cabau, C., Nguyen, T., Jouanno, E., Severac, D., Braasch, I., ... & Guiguen, Y. (2016). Gene evolution and gene expression after whole genome duplication in fish: the PhyloFish database. *BMC Genomics, 17*, 368.
- Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., ... & Reich, D. (2012). Ancient admixture in human history. *Genetics, 192*(3), 1065-1093.

- Pavey, S. A., Collin, H., Nosil, P., & Rogers, S. M. (2010). The role of gene expression in ecological speciation. *Annals of the New York Academy of Sciences*, 1206(1), 110.
- Pavlidis, P., & Alachiotis, N. (2017). A survey of methods and tools to detect recent and strong positive selection. *Journal of Biological Research-Thessaloniki*, 24(1), 7.
- Pavlidis, P., Živković, D., Stamatakis, A., & Alachiotis, N. (2013). SweeD: likelihood-based detection of selective sweeps in thousands of genomes. *Molecular Biology and Evolution*, 30(9), 2224-2234.
- Pedersen, S. H., Ferchaud, A. L., Bertelsen, M. S., Bekkevold, D., & Hansen, M. M. (2017). Low genetic and phenotypic divergence in a contact zone between freshwater and marine sticklebacks: gene flow constrains adaptation. *BMC Evolutionary Biology*, 17(1), 1-16.
- Perreault-Payette, A., Muir, A. M., Goetz, F., Perrier, C., Normandeau, E., Sirois, P., & Bernatchez, L. (2017). Investigating the extent of parallelism in morphological and genomic divergence among lake trout ecotypes in Lake Superior. *Molecular Ecology*, 26, 1477-1497.
- Pesquita, C., Faria, D., Falcao, A. O., Lord, P., & Couto, F. M. (2009). Semantic similarity in biomedical ontologies. *PLoS Computational Biology*, 5, e1000443.
- Pettengill, J. B., & Moeller, D. A. (2012). Phylogeography of speciation: allopatric divergence and secondary contact between outcrossing and selfing *Clarkia*. *Molecular Ecology*, 21(18), 4578-4592.
- Peyrégne, S., Boyle, M. J., Dannemann, M., & Prüfer, K. (2017). Detecting ancient positive selection in humans using extended lineage sorting. *Genome Research*, 27(9), 1563-1572.
- Phillips, R. B., Keatley, K. A., Morasch, M. R., Ventura, A. B., Lubieniecki, K. P., Koop, B.F., ... & Davidson, W. S. (2009). Assignment of Atlantic salmon (*Salmo salar*) linkage groups to specific chromosomes: conservation of large syntenic blocks corresponding to whole chromosome arms in rainbow trout (*Oncorhynchus mykiss*). *BMC Genetics*, 10, 46.

- 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(1), 196-205.
- Poisot, T., Bever, J. D., Nemri, A., Thrall, P. H., & Hochberg, M. E. (2011). A conceptual framework for the evolution of ecological specialisation. *Ecology Letters*, 14(9), 841-851.
- Pond, S. L. K., & Muse, S. V. (2005). HyPhy: hypothesis testing using phylogenies. In *Statistical Methods in Molecular Evolution* (pp. 125-181). Springer, New York, NY.
- Popescu, A. A., Huber, K. T., & Paradis, E. (2012). ape 3.0: New tools for distance-based phylogenetics and evolutionary analysis in R. *Bioinformatics*, 28, 1536-1537.
- Poulin, R. (2016). Greater diversification of freshwater than marine parasites of fish. *International Journal for Parasitology*, 46, 275-279.
- Primmer, C. R. (2011). Genetics of local adaptation in salmonid fishes. *Heredity*, 106, 401.
- Pritchard, J. K., Pickrell, J. K., & Coop, G. (2010). The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. *Current Biology*, 20(4), R208-R215.
- Pritchard, V. L., Mäkinen, H., Vähä, J. P., Erkinaro, J., Orell, P., & Primmer, C. R. (2018). Genomic signatures of fine-scale local selection in Atlantic salmon suggest involvement of sexual maturation, energy homeostasis and immune defence-related genes. *Molecular Ecology*, 27(11), 2560-2575.
- Puzey, J. R., Willis, J. H., & Kelly, J. K. (2017). Population structure and local selection yield high genomic variation in *Mimulus guttatus*. *Molecular Ecology*, 26(2), 519-535.

- Quevillon, E., Silventoinen, V., Pillai, S., Harte, N., Mulder, N., Apweiler, R., & Lopez, R. (2005). InterProScan: protein domains identifier. *Nucleic Acids Research*, 33, W116-W120.
- R Core Team (2016). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>
- R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: <https://www.R-project.org/>.
- Rabosky, D. L. (2013). Diversity-dependence, ecological speciation, and the role of competition in macroevolution. *Annual Review of Ecology, Evolution, and Systematics*, 44, 481-502.
- Racimo, F. (2016). Testing for ancient selection using cross-population allele frequency differentiation. *Genetics*, 202(2), 733-750.
- Racimo, F., Marnetto, D., & Huerta-Sánchez, E. (2017). Signatures of archaic adaptive introgression in present-day human populations. *Molecular Biology and Evolution*, 34(2), 296-317.
- Rahi, M. L., Amin, S., Mather, P. B., & Hurwood, D. A. (2017). Candidate genes that have facilitated freshwater adaptation by palaemonid prawns in the genus *Macrobrachium*: identification and expression validation in a model species (*M. koombooloomba*). *PeerJ*, 5, e2977.
- Ran, F. A., Hsu, P. D., Wright, J., Agarwala, V., Scott, D. A., & Zhang, F. (2013). Genome engineering using the CRISPR-Cas9 system. *Nature Protocols*, 8(11), 2281-2308.
- Rathjen, T., Yan, X., Kononenko, N. L., Ku, M. C., Song, K., Ferrarese, L., ... & Varela, L. (2017). Regulation of body weight and energy homeostasis by neuronal cell adhesion molecule 1. *Nature Neuroscience*, 20(8), 1096-1103.

- Rautsaw, R. M., Schramer, T. D., Acuña, R., Arick, L. N., DiMeo, M., Mercier, K. P., ... & Parkinson, C. L. (2020). Genomic Adaptations to Salinity Resist Gene Flow in the Evolution of Floridian Watersnakes. *Molecular Biology and Evolution*, msaa266.
- Rauwolf, U., Greiner, S., Mráček, J., Rauwolf, M., Golczyk, H., Mohler, V., ... & Meurer, J. (2011). Uncoupling of sexual reproduction from homologous recombination in homozygous *Oenothera* species. *Heredity*, 107(1), 87-94.
- Ravinet, M., Faria, R., Butlin, R. K., Galindo, J., Bierne, N., Rafajlović, M., ... & Westram, A. M. (2017). Interpreting the genomic landscape of speciation: a road map for finding barriers to gene flow. *Journal of Evolutionary Biology*, 30(8), 1450-1477.
- Ravinet, M., Yoshida, K., Shigenobu, S., Toyoda, A., Fujiyama, A., & Kitano, J. (2018). The genomic landscape at a late stage of stickleback speciation: High genomic divergence interspersed by small localized regions of introgression. *PLoS Genetics*, 14(5), e1007358.
- Recknagel, H., Elmer, K. R., & Meyer, A. (2014). Crater lake habitat predicts morphological diversity in adaptive radiations of cichlid fishes. *Evolution*, 68, 2145-2155.
- Recknagel, H., Hooker, O. E., Adams, C. E., & Elmer, K. R. (2017). Ecosystem size predicts eco-morphological variability in a postglacial diversification. *Ecology and Evolution*, 7, 5560-5570.
- Reeve, R., T. Leinster, C. Cobbold, J. Thompson, N. Brummitt, S. Mitchell, and L. Matthews. 2016. How to partition diversity. *arXiv*, 1404.6520:1-9.
- Refoyo-Martínez, A., da Fonseca, R. R., Halldórsdóttir, K., Árnason, E., Mailund, T., & Racimo, F. (2019). Identifying loci under positive selection in complex population histories. *Genome Research*, 29(9), 1506-1520.
- Reimand, J., Wagih, O., & Bader, G. D. (2015). Evolutionary constraint and disease associations of post-translational modification sites in human genomes. *PLoS Genetics*, 11, e1004919.

- Renaut, S., Grassa, C. J., Yeaman, S., Moyers, B. T., Lai, Z., Kane, N. C., ... & Rieseberg, L. H. (2013). Genomic islands of divergence are not affected by geography of speciation in sunflowers. *Nature Communications*, 4(1), 1-8.
- Rennison, D. J., Stuart, Y. E., Bolnick, D. I., & Peichel, C. L. (2019). Ecological factors and morphological traits are associated with repeated genomic differentiation between lake and stream stickleback. *Philosophical Transactions of the Royal Society B*, 374(1777), 20180241.
- Rettelbach, A., Nater, A., & Ellegren, H. (2019). How linked selection shapes the diversity landscape in *Ficedula* flycatchers. *Genetics*, 212(1), 277-285.
- Revelle, W. (2020). Psych: Procedures for Psychological, Psychometric, and Personality Research. R package version 1.9. 12.31.
- Revelle, W., & Revelle, M. W. (2015). Package ‘psych’. *The comprehensive R archive network*.
- Richards, E. J., & Martin, C. H. (2017). Adaptive introgression from distant Caribbean islands contributed to the diversification of a microendemic adaptive radiation of trophic specialist pupfishes. *PLoS Genetics*, 13(8), e1006919.
- Richards, E. J., Poelstra, J. W., & Martin, C. H. (2018). Don't throw out the sympatric speciation with the crater lake water: fine-scale investigation of introgression provides equivocal support for causal role of secondary gene flow in one of the clearest examples of sympatric speciation. *Evolution Letters*, 2(5), 524-540.
- Roberts, L. (2019). Signatures of selection in the red fox (*Vulpes vulpes*) in Europe using genotype-by-sequencing (Doctoral dissertation, University of Salford).
- Robertson, F. M., Gundappa, M. K., Grammes, F., Hvidsten, T. R., Redmond, A. K., Lien, S., Martin, S. A. M., Holland, P. W. H., Sandve, S. R., & Macqueen, D. J. (2017). Lineage-specific rediploidization is a mechanism to explain time-lags between genome duplication and evolutionary diversification. *Genome Biology*, 18, 111.

- Rodríguez-Ramilo, S. T., Baranski, M., Moghadam, H., Grove, H., Lien, S., Goddard, M. E., ... & Sonesson, A. K. (2019). Strong selection pressures maintain divergence on genomic islands in Atlantic cod (*Gadus morhua* L.) populations. *Genetics Selection Evolution*, 51(1), 1-15.
- Roesti, M., Gavrillets, S., Hendry, A. P., Salzburger, W., & Berner, D. (2014). The genomic signature of parallel adaptation from shared genetic variation. *Molecular Ecology*, 23(16), 3944-3956.
- Rondeau, E. B., Minkley, D. R., Leong, J. S., Messmer, A. M., Jantzen, J. R., von Schalburg, K. R., ... & Koop, B. F. (2014). The genome and linkage map of the northern pike (*Esox lucius*): conserved synteny revealed between the salmonid sister group and the Neoteleostei. *PLoS ONE*, 9, e102089.
- Rosado, A., & Raikhel, N. V. (2010). Application of the gene dosage balance hypothesis to auxin-related ribosomal mutants in Arabidopsis. *Plant Signaling & Behavior*, 5, 450-452.
- Rottner, K., & Stradal, T. E. (2011). Actin dynamics and turnover in cell motility. *Current Opinion in Cell Biology*, 23, 569-578.
- Rougeux, C., Bernatchez, L., & Gagnaire, P. A. (2017). Modeling the multiple facets of speciation-with-gene-flow toward inferring the divergence history of lake whitefish species pairs (*Coregonus clupeaformis*). *Genome Biology and Evolution*, 9(8), 2057-2074.
- Rougeux, C., Gagnaire, P. A., Praebel, K., Seehausen, O., & Bernatchez, L. (2019). Polygenic selection drives the evolution of convergent transcriptomic landscapes across continents within a Nearctic sister species complex. *Molecular Ecology*, 28(19), 4388-4403.
- Roux, J., Liu, J., & Robinson-Rechavi, M. (2017). Selective constraints on coding sequences of nervous system genes are a major determinant of duplicate gene retention in vertebrates. *Molecular Biology and Evolution*, 34, 2773-2791.

- Roux, J., Privman, E., Moretti, S., Daub, J. T., Robinson-Rechavi, M., & Keller, L. (2014). Patterns of positive selection in seven ant genomes. *Molecular Biology and Evolution*, 31, 1661-1685.
- Rui, L. (2013). Brain regulation of energy balance and body weight. *Reviews in Endocrine and Metabolic Disorders*, 14(4), 387-407.
- Rundle, H. D., & Nosil, P. (2005). Ecological speciation. *Ecology Letters*, 8(3), 336-352.
- Rustagi, N., Zhou, A., Watkins, W. S., Gedvilaite, E., Wang, S., Ramesh, N., ... & Xing, J. (2017). Extremely low-coverage whole genome sequencing in South Asians captures population genomics information. *BMC Genomics*, 18(1), 1-12.
- Rutigliano, G., Accorroni, A., & Zucchi, R. (2018). The case for TAAR1 as a modulator of central nervous system function. *Frontiers in Pharmacology*, 8, 987.
- Sabeti, P. C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., ... & Schaffner, S. F. (2007). Genome-wide detection and characterization of positive selection in human populations. *Nature*, 449(7164), 913-918.
- Saftig, P., Hunziker, E., Wehmeyer, O., Jones, S., Boyde, A., Rommerskirch, W., ... & Von Figura, K. (1998). Impaired osteoclastic bone resorption leads to osteopetrosis in cathepsin-K-deficient mice. *Proceedings of the National Academy of Sciences*, 95, 13453-13458.
- Salzburger, W. (2018). Understanding explosive diversification through cichlid fish genomics. *Nature Reviews Genetics*, 19(11), 705-717.
- Sandve, S. R., Rohlfs, R. V., Hvidsten, T. R. (2018). Subfunctionalization versus neofunctionalization after whole-genome duplication. *Nature Genetics*, 50, 908-909.
- Sano, S., Takashima, S., Niwa, H., Yokoi, H., Shimada, A., Arenz, A., ... & Takeda, H. (2009). Characterization of teleost Mdga1 using a gene-trap approach in medaka (*Oryzias latipes*). *Genesis*, 47(8), 505-513.

- Santini, F., Harmon, L. J., Carnevale, G., & Alfaro, M. E. (2009). Did genome duplication drive the origin of teleosts? A comparative study of diversification in ray-finned fishes. *BMC Evolutionary Biology*, 9, 194.
- Santos, M. E., Braasch, I., Boileau, N., Meyer, B. S., Sauteur, L., Böhne, A., ... & Salzburger, W. (2014). The evolution of cichlid fish egg-spots is linked with a cis-regulatory change. *Nature Communications*, 5, 5149.
- Savolainen, V., Anstett, M. C., Lexer, C., Hutton, I., Clarkson, J. J., Norup, M. V., ... & Baker, W. J. (2006). Sympatric speciation in palms on an oceanic island. *Nature*, 441(7090), 210-213.
- Schadt, E. E., Linderman, M. D., Sorenson, J., Lee, L., & Nolan, G. P. (2010). Computational solutions to large-scale data management and analysis. *Nature Reviews Genetics*, 11(9), 647-657.
- Schlamp, F., van der Made, J., Stambler, R., Chesebrough, L., Boyko, A. R., & Messer, P. W. (2016). Evaluating the performance of selection scans to detect selective sweeps in domestic dogs. *Molecular Ecology*, 25(1), 342-356.
- Schlötterer, C., Kofler, R., Versace, E., Tobler, R., & Franssen, S. U. (2015). Combining experimental evolution with next-generation sequencing: a powerful tool to study adaptation from standing genetic variation. *Heredity*, 114(5), 431-440.
- Schluter, D. (2016). Speciation, Ecological Opportunity, and Latitude: (American Society of Naturalists Address). *The American Naturalist*, 187, 1-18.
- Schluter, D., & Conte, G. L. (2009). Genetics and ecological speciation. *Proceedings of the National Academy of Sciences*, 106, 9955-9962.
- Schluter, D. (1996). Ecological causes of adaptive radiation. *The American Naturalist*, 148, S40-S64.
- Schluter, D. (1996). Ecological speciation in postglacial fishes. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 351(1341), 807-814.

- Schluter, D., & McPhail, J. D. (1992). Ecological character displacement and speciation in sticklebacks. *The American Naturalist*, 140(1), 85-108.
- Schneider, K., Adams, C. E., & Elmer, K. R. (2019). Parallel selection on ecologically relevant gene functions in the transcriptomes of highly diversifying salmonids. *BMC Genomics*, 20(1), 1-23.
- Schneider, K., White, T. J., Mitchell, S., Adams, C. E., Reeve, R., & Elmer, K. R. (2020). Pitfalls and virtues of population genetic summary statistics: detecting selective sweeps in recent divergences. *Journal of Evolutionary Biology*, DOI: <https://doi.org/10.1111/jeb.13738>.
- Schrider, D. R., & Kern, A. D. (2016). S/HIC: robust identification of soft and hard sweeps using machine learning. *PLoS Genetics*, 12(3).
- Schrider, D. R., & Kern, A. D. (2017). Soft sweeps are the dominant mode of adaptation in the human genome. *Molecular Biology and Evolution*, 34(8), 1863-1877.
- Schrider, D. R., & Kern, A. D. (2018). Supervised machine learning for population genetics: a new paradigm. *Trends in Genetics*, 34(4), 301-312.
- Schuler, H., Hood, G. R., Egan, S. P., & Feder, J. L. (2006). Modes and mechanisms of speciation. *Reviews in Cell Biology and Molecular Medicine*, 2(3), 60-93.
- Seehausen, O., Mayhew, P. J., & Van Alphen, J. J. M. (1999). Evolution of colour patterns in East African cichlid fish. *Journal of Evolutionary Biology*, 12(3), 514-534.
- Shakhnovich, B. E., & Koonin, E. V. (2006). Origins and impact of constraints in evolution of gene families. *Genome Research*, 16, 1529-1536.
- Sharma, V., Hecker, N., Roscito, J. G., Foerster, L., Langer, B. E., & Hiller, M. (2018). A genomics approach reveals insights into the importance of gene losses for mammalian adaptations. *Nature Communications*, 9, 1215.

- Shedko, S. V. (2019). Assembly ASM291031v2 (Genbank: GCA_002910315. 2) identified as assembly of the Northern Dolly Varden (*Salvelinus malma malma*) genome, and not the Arctic char (*S. alpinus*) genome. *arXiv*:1912.02474.
- Sheehan, S., & Song, Y. S. (2016). Deep learning for population genetic inference. *PLoS Computational Biology*, 12(3), e1004845.
- Shinde, S. S., Teekas, L., Sharma, S., & Vijay, N. (2019). Signatures of relaxed selection in the CYP8B1 gene of birds and mammals. *Journal of Molecular Evolution*, 87, 209-220.
- Signor, S. A., & Nuzhdin, S. V. (2018). The Evolution of Gene Expression in cis and trans. *Trends in Genetics*, 34(7), 532-544.
- Sikkink, K. L., & Snell-Rood, E. C. (2016). Plasticity paves the way in an adaptive radiation. *Molecular Ecology*, 25(24), 6009-6011.
- Singh, P., Börger, C., More, H., & Sturmbauer, C. (2017). The role of alternative splicing and differential gene expression in cichlid adaptive radiation. *Genome Biology and Evolution*, 9(10), 2764-2781.
- Siwertsson, A., Knudsen, R., Kahilainen, K., Præbel, K., Primicerio, R., & Amundsen, P. A. (2010). Sympatric diversification as influenced by ecological opportunity and historical contingency in a young species lineage of whitefish. *Evolutionary Ecology Research*, 12, 929-947.
- Siwertsson, A., Refsnes, B., Frainer, A., Amundsen, P. A., & Knudsen, R. (2016). Divergence and parallelism of parasite infections in Arctic charr morphs from deep and shallow lake habitats. *Hydrobiologia*, 783, 131-143.
- Škaloud, P., Škaloudová, M., Doskočilová, P., Kim, J. I., Shin, W., & Dvořák, P. (2019). Speciation in protists: Spatial and ecological divergence processes cause rapid species diversification in a freshwater chrysophyte. *Molecular Ecology*, DOI: <https://doi.org/10.1111/mec.15011>.

- Skotte, L., Korneliussen, T. S., & Albrechtsen, A. (2013). Estimating individual admixture proportions from next generation sequencing data. *Genetics*, 195(3), 693-702.
- Skúlason, S., Snorrason, S. S., Noakes, D. L., & Ferguson, M. M. (1996). Genetic basis of life history variations among sympatric morphs of Arctic char *Salvelinus alpinus*. *Canadian Journal of Fisheries and Aquatic Sciences*, 53(8), 1807-1813.
- Smith, M. D., Wertheim, J. O., Weaver, S., Murrell, B., Scheffler, K., & Kosakovsky Pond, S. L. (2015). Less is more: an adaptive branch-site random effects model for efficient detection of episodic diversifying selection. *Molecular Biology and Evolution*, 32(5), 1342-1353.
- Solbakken, M. H., Voje, K. L., Jakobsen, K. S., & Jentoft, S. (2017). Linking species habitat and past palaeoclimatic events to evolution of the teleost innate immune system. *Proceedings of the Royal Society B: Biological Sciences*, 284, 20162810.
- Solomon, K. S., Kudoh, T., Dawid, I. B., & Fritz, A. (2003). Zebrafish foxi1 mediates otic placode formation and jaw development. *Development*, 130, 929-940.
- Soraggi, S., Wiuf, C., & Albrechtsen, A. (2018). Powerful inference with the D-statistic on low-coverage whole-genome data. *G3: Genes, Genomes, Genetics*, 8(2), 551-566.
- Stajich, J. E., & Hahn, M. W. (2005). Disentangling the effects of demography and selection in human history. *Molecular Biology and Evolution*, 22(1), 63-73.
- Stankowski, S., Chase, M. A., Fuiten, A. M., Rodrigues, M. F., Ralph, P. L., & Streisfeld, M. A. (2019). Widespread selection and gene flow shape the genomic landscape during a radiation of monkeyflowers. *PLoS Biology*, 17(7), e3000391.
- Stapley, J., Feulner, P. G., Johnston, S. E., Santure, A. W., & Smadja, C. M. (2017). Variation in recombination frequency and distribution across eukaryotes: patterns and processes. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 372(1736), 20160455.
- Stearns, S. C., & Hendry, A. P. (2004). The salmonid contribution to key issues in evolution. *Evolution illuminated: Salmon and their relatives*, 3-19.

- Stecher, G., Tamura, K., & Kumar, S. (2020). Molecular evolutionary genetics analysis (MEGA) for macOS. *Molecular Biology and Evolution*, 37(4), 1237-1239.
- Steinhäuser, S. S. (2013). Characterization of natterin-like genes in Arctic charr (*Salvelinus alpinus*). (Doctoral dissertation). University of Iceland, Reykjavík, Iceland.
- Stern, D. L., & Orgogozo, V. (2009). Is genetic evolution predictable?. *Science*, 323(5915), 746-751.
- Stinchcombe, J. R., & Hoekstra, H. E. (2008). Combining population genomics and quantitative genetics: finding the genes underlying ecologically important traits. *Heredity*, 100(2), 158-170.
- Streisfeld, M. A., & Kohn, J. R. (2005). Contrasting patterns of floral and molecular variation across a cline in *Mimulus aurantiacus*. *Evolution*, 59(12), 2548-2559.
- Striberny, A., Jørgensen, E. H., Klopp, C., & Magnanou, E. (2019). Arctic charr brain transcriptome strongly affected by summer seasonal growth but only subtly by feed deprivation. *BMC Genomics*, 20, 529.
- Stroud, J. T., & Losos, J. B. (2016). Ecological opportunity and adaptive radiation. *Annual Review of Ecology, Evolution, and Systematics*, 47, 507-532.
- Stuart, Y. E., Veen, T., Weber, J. N., Hanson, D., Ravinet, M., Lohman, B. K., ... & Ahmed, N. (2017). Contrasting effects of environment and genetics generate a continuum of parallel evolution. *Nature Ecology & Evolution*, 1(6), 1-7.
- Stutz, W. E., & Bolnick, D. I. (2017). Natural selection on MHC II β in parapatric lake and stream stickleback: Balancing, divergent, both or neither?. *Molecular Ecology*, 26(18), 4772-4786.
- Su, S., Li, H., Du, F., Zhang, C., Li, X., Jing, X., ... & Yuan, X. (2018). Combined QTL and genome scan analyses with the help of 2b-RAD identify growth-associated genetic markers in a new fast-growing carp strain. *Frontiers in Genetics*, 9, 592.

- Supek, F., Bošnjak, M., Škunca, N., & Šmuc, T. (2011). REVIGO summarizes and visualizes long lists of gene ontology terms. *PloS ONE*, 6, e21800.
- Suzuki, Y., & Nei, M. (2004). False-positive selection identified by ML-based methods: examples from the Sig1 gene of the diatom *Thalassiosira weissflogii* and the tax gene of a human T-cell lymphotropic virus. *Molecular Biology and Evolution*, 21, 914-921.
- Svardal, H., Quah, F. X., Malinsky, M., Ngatunga, B. P., Miska, E. A., Salzburger, W., ... & Durbin, R. (2020). Ancestral hybridization facilitated species diversification in the Lake Malawi cichlid fish adaptive radiation. *Molecular Biology and Evolution*, 37(4), 1100-1113.
- Sweigart, A. L., Brandvain, Y., & Fishman, L. (2019). Making a murderer: the evolutionary framing of hybrid gamete-killers. *Trends in Genetics*, 35(4), 245-252.
- Swings, T., Van den Bergh, B., Wuyts, S., Oeyen, E., Voordeckers, K., Verstrepen, K. J., ... & Michiels, J. (2017). Adaptive tuning of mutation rates allows fast response to lethal stress in *Escherichia coli*. *Elife*, 6, e22939.
- Tajima, F. (1983). Evolutionary relationship of DNA sequences in finite populations. *Genetics*, 105(2), 437-460.
- Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, 123(3), 585-595.
- Takahashi, D., Hase, K., Kimura, S., Nakatsu, F., Ohmae, M., Mandai, Y., ... & Obata, Y. (2011). The epithelia-specific membrane trafficking factor AP-1B controls gut immune homeostasis in mice. *Gastroenterology*, 141, 621-632.
- Takahata, N., & Nei, M. (1985). Gene genealogy and variance of interpopulational nucleotide differences. *Genetics*, 110(2), 325-344.
- Tan, K., An, L., Wang, S. M., Wang, X. D., Zhang, Z. N., Miao, K., ... & Tian, J. H. (2015). Actin disorganization plays a vital role in impaired embryonic development of in vitro-produced mouse preimplantation embryos. *PloS ONE*, 10, e0130382.

- Tang, H., & Thomas, P. D. (2016). Tools for predicting the functional impact of nonsynonymous genetic variation. *Genetics*, 203(2), 635-647.
- Taylor, J. S., Van de Peer, Y., & Meyer, A. (2001). Genome duplication, divergent resolution and speciation. *Trends in Genetics*, 17, 299-301.
- Tebich, S., Sterelny, K., & Teschke, I. (2010). The tale of the finch: adaptive radiation and behavioural flexibility. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1543), 1099-1109.
- Tessarolo, J. A., Tabesh, M. J., Nesbitt, M., & Davidson, W. S. (2014). Genomic organization and evolution of the trace amine-associated receptor (TAAR) repertoire in Atlantic salmon (*Salmo salar*). *G3: Genes, Genomes, Genetics*, 4(6), 1135-1141.
- Therkildsen, N. O., & Palumbi, S. R. (2017). Practical low-coverage genomewide sequencing of hundreds of individually barcoded samples for population and evolutionary genomics in nonmodel species. *Molecular Ecology Resources*, 17(2), 194-208.
- Thornton, K. (2005). Recombination and the properties of Tajima's D in the context of approximate-likelihood calculation. *Genetics*, 171(4), 2143-2148.
- Thurman, T. J., & Barrett, R. D. (2016). The genetic consequences of selection in natural populations. *Molecular Ecology*, 25(7), 1429-1448.
- To, T. T., Witten, P. E., Huysseune, A., & Winkler, C. (2015). An adult osteopetrosis model in medaka reveals the importance of osteoclast function for bone remodeling in teleost fish. *Comparative Biochemistry and Physiology Part C: Toxicology & Pharmacology*, 178, 68-75.
- Tomalty, K. M., Meek, M. H., Stephens, M. R., Rincón, G., Fangue, N. A., May, B. P., & Baerwald, M. R. (2015). Transcriptional response to acute thermal exposure in juvenile chinook salmon determined by RNAseq. *G3: Genes, Genomes, Genetics*, 5, 1335-1349.

- Torres-Dowdall, J., Henning, F., Elmer, K. R., & Meyer, A. (2015). Ecological and lineage-specific factors drive the molecular evolution of rhodopsin in cichlid fishes. *Molecular Biology and Evolution*, 32(11), 2876-2882.
- Torres-Dowdall, J., Pierotti, M. E., Härer, A., Karagic, N., Woltering, J. M., Henning, F., ... & Meyer, A. (2017). Rapid and parallel adaptive evolution of the visual system of Neotropical Midas cichlid fishes. *Molecular Biology and Evolution*, 34(10), 2469-2485.
- Turner, J. R., & T, J. R. G. (1977). Butterfly mimicry: the genetical evolution of an adaptation. *Evolutionary Biology*, 10, 163-206.
- Urban, S., Nater, A., Meyer, A., & Kratochwil, C. F. (2020). Different sources of allelic variation drove repeated color pattern divergence in cichlid fishes. *Molecular Biology and Evolution*, msaa237.
- Varadharajan, S., Sandve, S. R., Gillard, G. B., Tørresen, O. K., Mulugeta, T. D., Hvidsten, T. R., ... & Jakobsen, K. S. (2018). The Grayling Genome Reveals Selection on Gene Expression Regulation after Whole-Genome Duplication. *Genome Biology and Evolution*, 10, 2785-2800.
- Vatsiou, A. I., Bazin, E., & Gaggiotti, O. E. (2016). Detection of selective sweeps in structured populations: a comparison of recent methods. *Molecular Ecology*, 25(1), 89-103.
- Velotta, J. P., McCormick, S. D., O'Neill, R. J., & Schultz, E. T. (2014). Relaxed selection causes microevolution of seawater osmoregulation and gene expression in landlocked Alewives. *Oecologia*, 175, 1081-1092.
- Velotta, J. P., Wegrzyn, J. L., Ginzburg, S., Kang, L., Czesny, S., O'Neill, R. J., ... & Schultz, E. T. (2017). Transcriptomic imprints of adaptation to fresh water: parallel evolution of osmoregulatory gene expression in the Alewife. *Molecular Ecology*, 26, 831-848.

- Verity, R., Collins, C., Card, D. C., Schaal, S. M., Wang, L., & Lotterhos, K. E. (2017). MINOTAUR: A platform for the analysis and visualization of multivariate results from genome scans with R Shiny. *Molecular Ecology Resources*, 17(1), 33-43.
- Verspoor, E., Knox, D., Greer, R., & Hammar, J. (2010). Mitochondrial DNA variation in Arctic charr (*Salvelinus alpinus* (L.)) morphs from Loch Rannoch, Scotland: evidence for allopatric and peripatric divergence. *Hydrobiologia*, 650(1), 117-131.
- Verta, J. P., & Jones, F. C. (2019). Predominance of cis-regulatory changes in parallel expression divergence of sticklebacks. *eLife*, 8, e43785.
- Vitti, J. J., Grossman, S. R., & Sabeti, P. C. (2013). Detecting natural selection in genomic data. *Annual review of genetics*, 47, 97-120.
- Vonlanthen, P., Roy, D., Hudson, A. G., Largiadèr, C. R., Bittner, D., & Seehausen, O. (2009). Divergence along a steep ecological gradient in lake whitefish (*Coregonus* sp.). *Journal of Evolutionary Biology*, 22(3), 498-514.
- Wagner, C. E., Harmon, L. J., & Seehausen, O. (2012). Ecological opportunity and sexual selection together predict adaptive radiation. *Nature*, 487(7407), 366-369.
- Wainwright, D. K. (2019). Fish scales: morphology, evolution, and function (Doctoral dissertation, Harvard University).
- Wallace, A. R. (1871). Contributions to the theory of natural selection. Macmillan and Company, London.
- Walsh, J., Kovach, A. I., Olsen, B. J., Shriver, W. G., & Lovette, I. J. (2018). Bidirectional adaptive introgression between two ecologically divergent sparrow species. *Evolution*, 72(10), 2076-2089.
- Wang, J., Vasaikar, S., Shi, Z., Greer, M., & Zhang, B. (2017). WebGestalt 2017: a more comprehensive, powerful, flexible and interactive gene set enrichment analysis toolkit. *Nucleic Acids Research*, 45, W130-W137.

- Wang, J. (2012). On the measurements of genetic differentiation among populations. *Genetics Research*, 94(5), 275-289.
- Wang, J., Duncan, D., Shi, Z., & Zhang, B. (2013). WEB-based gene set analysis toolkit (WebGestalt): update 2013. *Nucleic Acids Research*, 41(W1), W77-W83.
- Wang, J., Vasaikar, S., Shi, Z., Greer, M., & Zhang, B. (2017). WebGestalt 2017: a more comprehensive, powerful, flexible and interactive gene set enrichment analysis toolkit. *Nucleic Acids Research*, 45(W1), W130-W137.
- Ward, B. J., & Van Oosterhout, C. (2016). HYBRIDCHECK: software for the rapid detection, visualization and dating of recombinant regions in genome sequence data. *Molecular Ecology Resources*, 16(2), 534-539.
- Watterson, G. A. (1975). On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology*, 7(2), 256-276.
- Weismann, A. (1868). Über die Berechtigung der Darwin'schen Theorie: ein akademischer Vortrag gehalten am 8. Juli 1868 in der Aula der Universität zu Freiburg im Breisgau. W. Engelmann, Leipzig.
- Weismann, A. (1872). Ueber den Einfluss der Isolirung auf die Artbildung. W. Engelmann, Leipzig.
- Wellborn, G. A., & Langerhans, R. B. (2015). Ecological opportunity and the adaptive diversification of lineages. *Ecology and Evolution*, 5, 176-195.
- Wellenreuther, M., & Bernatchez, L. (2018). Eco-evolutionary genomics of chromosomal inversions. *Trends in Ecology & Evolution*, 33(6), 427-440.
- Wertheim, J. O., Murrell, B., Smith, M. D., Kosakovsky Pond, S. L., & Scheffler, K. (2015). RELAX: detecting relaxed selection in a phylogenetic framework. *Molecular Biology and Evolution*, 32(3), 820-832.

- Westram, A. M., Galindo, J., Alm Rosenblad, M., Grahame, J. W., Panova, M., & Butlin, R. K. (2014). Do the same genes underlie parallel phenotypic divergence in different *Littorina saxatilis* populations?. *Molecular Ecology*, 23(18), 4603-4616.
- Westram, A. M., Panova, M., Galindo, J., & Butlin, R. K. (2016). Targeted resequencing reveals geographical patterns of differentiation for loci implicated in parallel evolution. *Molecular Ecology*, 25(13), 3169-3186.
- Westram, A. M., Rafajlović, M., Chaube, P., Faria, R., Larsson, T., Panova, M., ... & Butlin, R. (2018). Clines on the seashore: The genomic architecture underlying rapid divergence in the face of gene flow. *Evolution Letters*, 2(4), 297-309.
- White, T. J. (2019). Developing new measures of genetic diversity (Unpublished Master's thesis). University of Glasgow, Glasgow, UK.
- Whitlock, M. C. (2000). Fixation of new alleles and the extinction of small populations: drift load, beneficial alleles, and sexual selection. *Evolution*, 54(6), 1855-1861.
- Wickham, H. (2011). ggplot2. Wiley Interdisciplinary Reviews: *Computational Statistics*, 3, 180-185.
- Wilder, A. P., Palumbi, S. R., Conover, D. O., & Therkildsen, N. O. (2020). Footprints of local adaptation span hundreds of linked genes in the Atlantic silverside genome. *Evolution Letters*, 4(5), 430-443.
- Wolf, J. B., & Ellegren, H. (2017). Making sense of genomic islands of differentiation in light of speciation. *Nature Reviews Genetics*, 18(2), 87.
- Wright, S. (1949). The genetical structure of populations. *Annals of Eugenics*, 15(1), 323-354.
- Wund, M. A., Baker, J. A., Clancy, B., Golub, J. L., & Foster, S. A. (2008). A test of the “flexible stem” model of evolution: ancestral plasticity, genetic accommodation, and morphological divergence in the threespine stickleback radiation. *The American Naturalist*, 172(4), 449-462.

- Xing, Y., & Lee, C. (2006). Alternative splicing and RNA selection pressure—evolutionary consequences for eukaryotic genomes. *Nature Reviews Genetics*, 7, 499.
- Yamasaki, Y. Y., Kakioka, R., Takahashi, H., Toyoda, A., Nagano, A. J., Machida, Y., ... & Kitano, J. (2020). Genome-wide patterns of divergence and introgression after secondary contact between *Pungitius* sticklebacks. *Philosophical Transactions of the Royal Society B*, 375(1806), 20190548.
- Yan, J., Zhang, Y., Cheng, S., Kang, B., Peng, J., Zhang, X., ... & Zhang, S. (2017). Common genetic heterogeneity of human interleukin-37 leads to functional variance. *Cellular & Molecular Immunology*, 14, 783.
- Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, 24, 1586-1591.
- Yang, Z., & Nielsen, R. (2002). Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Molecular Biology and Evolution*, 19, 908-917.
- Yoder, J. B., Clancey, E., Des Roches, S., Eastman, J. M., Gentry, L., Godsoe, W., ... & Sarver, B. A. J. (2010). Ecological opportunity and the origin of adaptive radiations. *Journal of Evolutionary Biology*, 23(8), 1581-1596.
- York, R. A., Byrne, A., Abdilleh, K., Patil, C., Streelman, T., Finger, T. E., & Fernald, R. D. (2019). Behavioral evolution contributes to hindbrain diversification among Lake Malawi cichlid fish. *Scientific Reports*, 9(1), 1-9.
- Zdobnov, E. M., & Apweiler, R. (2001). InterProScan – an integration platform for the signature-recognition methods in InterPro. *Bioinformatics*, 17, 847-848.
- Zhang, J. (2004). Frequent false detection of positive selection by the likelihood method with branch-site models. *Molecular Biology and Evolution*, 21, 1332-1339.

- Zhang, J., Nielsen, R., & Yang, Z. (2005). Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Molecular Biology and Evolution*, 22, 2472-2479.
- Zhang, B., Kirov, S., & Snoddy, J. (2005). WebGestalt: an integrated system for exploring gene sets in various biological contexts. *Nucleic Acids Research*, 33(suppl_2), W741-W748.
- Zhao, H., Soufan, O., Xia, J., Tang, R., Li, L., & Li, D. (2019). Transcriptome and physiological analysis reveal alterations in muscle metabolisms and immune responses of grass carp (*Ctenopharyngodon idellus*) cultured at different stocking densities. *Aquaculture*, 503, 186-197.
- Zheng, X., Gogarten, S. M., Lawrence, M., Stilp, A., Conomos, M. P., Weir, B. S., ... & Levine, D. (2017). SeqArray—a storage-efficient high-performance data format for WGS variant calls. *Bioinformatics*, 33(15), 2251-2257.
- Zheng, X., Levine, D., Shen, J., Gogarten, S. M., Laurie, C., & Weir, B. S. (2012). A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics*, 28(24), 3326-3328.
- Zheng, Y., & Wiehe, T. (2019). Adaptation in structured populations and fuzzy boundaries between hard and soft sweeps. *PLoS Computational Biology*, 15(11).
- Zizioli, D., Forlanelli, E., Guarienti, M., Nicoli, S., Fanzani, A., Bresciani, R., ... & Schu, P. (2010). Characterization of the AP-1 μ 1A and μ 1B adaptins in zebrafish (*Danio rerio*). *Developmental Dynamics*, 239, 2404-2412.